UNIVERSIDAD AUTÓNOMA DE MADRID

FACULTAD DE FILOSOFÍA Y LETRAS

Departamento de Lingüística General, Lenguas Modernas, Lógica y Filosofía de la Ciencia, Tª de la Literatura y Literatura Comparada

"ANÁLISIS DE ERRORES DE APRENDIENTES DE FRANCÉS LENGUA EXTRANJERA (FLE) BASADO EN CORPUS ORALES"

Ana Valverde Mateos

Tesis doctoral dirigida por el Dr. Antonio Moreno Sandoval (UAM) y la Dra. Concepción Sanz Miguel (UCLM)

2012

El presente proyecto de tesis ha sido financiado gracias a una beca de Formación del Profesorado Universitario (FPU), concedida por el Ministerio de Educación, con referencia AP2007-00111.

Contenido

ÍNDICE DE TABLAS

ÍNDICE DE GRÁFICOS

AGRADECIMIENTOS

"On n'apprend pas sans faire d'erreurs et les erreurs servent à apprendre"

(Porquier, 1977: 28)

En el proceso de investigación y redacción de esta tesis, muchas son las personas que se han implicado, y que han conseguido que este arduo trabajo llegue a su fin sin tener que lamentar excesivos daños colaterales.

En primer lugar, me gustaría agradecer a mis directores de tesis, el Dr. Antonio Moreno Sandoval y la Dra. Concha Sanz Miguel, todo el conocimiento que han puesto a mi alcance a lo largo de estos años. Mucho de lo que sé se lo debo a ellos, y estoy segura de que sin sus buenos consejos y enseñanzas, esta tesis nunca hubiera visto la luz. A Antonio le debo todo lo que sé de Lingüística de Corpus y le agradezco su confianza, su empatía y su generosidad. A Concha le debo todo lo que sé sobre enseñanza y sobre Humanidades Digitales, y le agradezco tantas cosas que me extendería tanto como de costumbre... He de destacar su tesón, su optimismo, su entrega y las grandes oportunidades que me ha brindado. (Sin ti, la universidad nunca será lo mismo. Ella y nosotras te echaremos de menos).

En segundo lugar, tengo que agradecer todo el trabajo y las investigaciones previas de los miembros del Laboratorio de Lingüística Informática de la UAM, que han permitido que mi investigación tenga una de las mejores metodologías y ejemplos en los que apoyarse. De todos ellos, he de destacar a Leonardo, mi *gurú* de los corpus de aprendientes y un gran apoyo investigador en cualquier latitud y longitud.

Mención especial merecen también todos los aprendientes que generosamente han colaborado conmigo y ponen voz a esta investigación. Gracias a las EOI y a la Facultad de Letras, y a los profesores y departamentos que nos acogieron, representados por Antonia Ortiz, Ramón García Pradas, y Luis Felipe Casero Martín.

No todo en la vida es trabajo y por tanto, he de agradecer a todos mis amigos su ayuda, sus gestos de ánimo y su cariño. A todos aquellos que preguntaron por la tesis y esperaron a oír la respuesta, más o menos desesperada: ¡muchas gracias! Y entre ellos: Patricia, Maribel, Sara, Valeria, Celia, David, Carlos J., María, Álvaro, Marian, el informático de los comandos, y mis físicas *grenobloises,* Laura y Rocío.

Gracias a Carlos A. por tu disponibilidad y por las ayudas de último minuto que alivian trabajos titánicos.

Nada de esto hubiera sido posible sin Silvia, no sólo por tener en ella a mi mejor consultora informática, sino por ser mi apoyo día a día. Gracias por hacer que la universidad sea un segundo hogar y por las risas desde las nueve de la mañana. Albacete y la UCLM no hubieran sido lo mismo sin haberte conocido.

Por último, aunque por supuesto, no menos importante, he de agradecer mucho de lo que soy a mis padres y a mi abuela. Gracias por el apoyo moral (y también económico) a lo largo de estos años, por confiar en mí, y sobre todo, por aguantarme en las *neuras*. Es una suerte saber que siempre estáis ahí, tan cerca.

INTRODUCCIÓN

Hace quince o veinte años, la investigación sobre muestras colosales de datos legibles por los ordenadores, es decir *los corpus*, todavía estaba en manos de una avanzadilla muy minoritaria de estudiosos. En el momento actual, el aumento considerable del interés por la comprensión intelectual del lenguaje humano y por el desarrollo de lenguajes artificiales útiles para la ingeniería informática, la han convertido en un foco de atención muy importante para otras ciencias auxiliares que se benefician de sus progresos.

El presente proyecto de investigación lleva por título: "Análisis de errores de aprendientes de Francés como Lengua Extranjera (FLE) basado en corpus orales", y propone la aplicación de los corpus orales al ámbito de la enseñanza de lenguas, como medio de diagnóstico y descripción de la interlengua (o competencia de comunicación transitoria) de los aprendientes hispanohablantes de francés.

Su contenido se encuadra en la intersección de la Lingüística de Corpus, la metodología de enseñanza de las lenguas y la metodología tradicional para el análisis de datos. La técnica elegida es un sistema de análisis de errores basado en taxonomías existentes, utilizadas anteriormente para la confección de otros corpus orales de aprendientes de francés de ámbito europeo, como el corpus FRIDA (*FRench Interlanguage DAtabase)*, creado por el proyecto europeo FreeText, cuyo propósito era desarrollar un sistema de diagnóstico automático de errores para aprendientes de FLE.

Justificación del proyecto

En su barómetro de febrero 2010[1], el Centro de Investigaciones Sociológicas (CIS) nos recordaba que la mayor parte de los españoles no conocía ningún idioma extranjero, y que más del 90% no estudiaba ninguno. Ese mismo año, la oficina estadística europea, Eurostat, confirmaba lo que era un secreto a voces: casi un 46,6% de los españoles

[1] http://www.cis.es/cis/opencm/ES/1_encuestas/estudios/ver.jsp?estudio=10082

entre 25 y 64 años no tenía noción alguna de ningún idioma extranjero, frente al 38,3% de la media de la Unión Europea[2].

No es lugar aquí para indagar las razones profundas de la mala salud del aprendizaje de las lenguas en España, pero sí para dejar constancia de nuestra sospecha de que algo tendrán que ver la frustración y desmotivación que a menudo dicen experimentar los estudiantes cuando constatan que los hablantes nativos parecen manejar una lengua distinta a la que ellos han aprendido en clase. Muchos de ellos se quejan asimismo de sus excesivas dificultades para expresarse oralmente, teniendo en cuenta la cantidad de energía y de tiempo que invierten en el aprendizaje del habla.

El conjunto de estas reflexiones se encuentra en el origen de nuestro convencimiento de que una profundización en el conocimiento de la interlengua de los aprendientes hispanohablantes aprendientes de francés, mediante un análisis exhaustivo de los errores más frecuentes que cometen en las prácticas de interacción oral en contexto educativo, podrían contribuir a esclarecer algunas causas del tan comentado desfase entre esfuerzo desplegado y resultados obtenidos. Tal vez, y esa es nuestra hipótesis de trabajo, la información sobre los errores generalmente asociados a ciertos procedimientos psicolingüísticos subyacentes, nos ayude a tipificar las dificultades concretas de desarrollo de una comunicación oral efectiva para este público meta específico.

Docentes y estudiantes están emplazados a alcanzar los objetivos de comunicación promulgados por el Marco Común Europeo de Referencia (MCER), eje común para el desarrollo de los currícula del aprendizaje de las lenguas a nivel europeo. Aspiramos a que nuestro estudio contribuya a una mejor fundamentación de los enfoques pedagógicos, que ayude a los profesores a planificar su trabajo de manera más efectiva, y a los aprendientes a mejorar su competencia oral de forma progresiva, adecuándola a los distintos contextos, situaciones comunicativas y tipos de hablantes nativos.

Por otra parte, estamos también convencidos de que la observación de producciones orales auténticas sistematizadas y procedentes de entornos no nativos ofrecerá a los docentes

---

2 http://epp.eurostat.ec.europa.eu/portal/page/portal/product_details/dataset?p_product_code=EDUC_THFRLAN

(generalmente, también no nativos), una oportunidad de abordar la lengua de estudio de una manera diferente, más adaptada a las dificultades objetivas, y que presenta nuevas vías por explorar. Creemos firmemente que la didáctica de las lenguas no puede ser efectiva si no se apoya en la individualización y en la adaptación a las necesidades reales de comunicación de los alumnos. A estos últimos, trabajar con corpus les brindará la oportunidad de ensayar un nuevo paradigma de estudio.

Un análisis exhaustivo de una variedad de lengua, sólo puede darse de manera productiva a partir de una cantidad suficiente de datos. Dichos datos deben poseer, además, unas características técnicas muy específicas, que les permite ser fácilmente almacenados, consultados y reutilizados, mediante las distintas herramientas informáticas hoy disponibles. Nada de esto es posible sin la existencia de un corpus, en este caso de un corpus oral de aprendientes.

Pese a ello, estamos en condiciones de afirmar que no existen en la actualidad corpus orales de aprendientes hispanohablantes de FLE publicados o difundidos entre la comunidad científica. Los corpus que se conocen se reducen a corpus escritos muy específicos y a corpus basados en la interacción (asíncrona) a través del desarrollo de actividades de comunicación asistida por ordenador (CMC o *Computer Mediated Communication*).

El corpus que hemos creado se denomina CORAF (*Corpus ORal de Aprendientes de Francés*), y está diseñado e implementado mediante la metodología desarrollada para la compilación de corpus orales por el Laboratorio de Lingüística Informática de la Universidad Autónoma de Madrid.

CORAF podrá convertirse en la base para la implementación de futuros análisis de actuaciones o del discurso. En la presente tesis, daremos cuenta del desarrollo de la aplicación que suele ser prioritaria para este tipo de corpus: el análisis sistemático de errores. En este caso, caracterizaremos los errores desde tres puntos de vista: el lingüístico, el descriptivo y el etiológico, siendo este último el que más datos nos proporcione sobre los procesos que moviliza el aprendiente para la construcción de su interlengua.

Objetivos del proyecto

En resumidas cuentas, los objetivos específicos de nuestra investigación son los siguientes:

- Ampliación significativa del conocimiento disponible sobre la interlengua de los aprendientes hispanohablantes de FLE;
- Creación de un corpus oral de aprendientes de tipo pedagógico por medio de la recogida de muestras, para su posterior transcripción ortográfica, alineamiento y conversión a formatos estándar y reutilizables;
- Realización de un análisis sistemático de errores manual desde tres puntos de vista: lingüístico, descriptivo y etiológico;
- Extracción de datos del corpus mediante métodos automáticos y estadísticos;
- Análisis de los resultados obtenidos, con vistas a ofrecer una panorámica de las dificultades más habituales entre los aprendientes de FLE de todos los niveles del MCER cubiertos por nuestro corpus.

El Corpus CORAF y el análisis de errores conforman una base de datos para la caracterización del aprendiente, de una riqueza poco frecuente en este tipo de trabajos. El recurso estará destinado al conjunto de la comunidad investigadora, que podrá utilizarlo de forma gratuita para la realización de nuevos estudios y para la implementación de herramientas o aplicaciones destinadas al aprendizaje del francés.

Estructura del estudio

El presente texto se divide en dos partes: la primera engloba el marco teórico y metodológico en el que se inscribe el trabajo, y la segunda describe la aplicación propiamente dicha que constituye el objeto principal de nuestro estudio: un corpus oral de aprendientes de

FLE y un análisis de los errores más frecuentes entre la cohorte de aprendientes que participaron en el estudio.

La primera parte consta de seis capítulos. El primero está dedicado a la descripción, caracterización y distinción entre la Lingüística Computacional y la Lingüística de Corpus, y explica los enfoques teóricos y metodológicos utilizados para nuestra investigación.

El segundo capítulo se ocupa de definir el concepto de corpus, y describir sus características y usos, ofreciendo una panorámica de la Lingüística de Corpus en Francia, y realizando un breve recorrido por los corpus más representativos del francés, por ser esta la lengua objeto de nuestro estudio.

El tercer capítulo pretende mostrar la aplicación de los corpus a la enseñanza y, más concretamente, a la enseñanza de las lenguas, analizando los proyectos y enfoques más utilizados en su aplicación (directa, indirecta y para la elaboración de corpus pedagógicos), tanto en España como en Francia.

Posteriormente, el cuarto capítulo aborda el concepto de sujeto de aprendizaje o aprendiente, deteniéndose en los factores externos e internos que pueden influir en el proceso de adquisición de la lengua y, más concretamente, en el uso de los corpus como herramienta de aprendizaje.

El quinto capítulo examina las relaciones eventuales entre los corpus y los aprendientes de lenguas, configurando las características, usos y aplicaciones más frecuentes, así como efectuando un breve recorrido por los corpus de aprendientes del francés existentes en el mercado.

Por último, el sexto capítulo aborda la metodología de análisis de datos que se utiliza con más frecuencia con los corpus. Se recorre brevemente la historia y evolución del Análisis de Errores, describiendo los enfoques metodológicos empleados hasta el momento y destacando las taxonomías a las que han dado lugar. Esta descripción servirá, además, para establecer las taxonomías elegidas para nuestro propio sistema de análisis de errores.

La segunda parte de la tesis se centra en el objetivo operativo de nuestro estudio: el análisis de errores de aprendientes de francés en situación de interacción oral, partiendo del corpus oral implementado.

Esta segunda parte se compone de cuatro capítulos. En el primero, se describe el diseño y la concepción del corpus oral de aprendientes CORAF.

El segundo capítulo muestra el proceso de compilación e implementación del corpus, describiendo todas las fases que hemos seguido para su realización, y deteniéndonos en algunos de los procesos más importantes, como la recogida de datos y la transcripción del sonido.

El tercero muestra una de las posibles aplicaciones del corpus oral implementado, que es el análisis de errores frecuentes en aprendientes. Así, se muestran los resultados obtenidos para el corpus CORAF por medio de los criterios de análisis seleccionados y centrándonos en los seis niveles de dominio del MCER que cubre nuestro estudio.

La segunda parte finaliza con el cuarto capítulo que recoge las conclusiones finales del proyecto y describe los posibles ejes futuros de prolongación de la investigación.

El trabajo contiene, además, dos apéndices y un glosario de siglas y abreviaturas. El primer apéndice presenta de forma más detallada varios ejemplos de transcripción (apéndice A), y el segundo muestra el conjunto de listas de frecuencias de errores categorizados teniendo en cuenta todos los criterios estudiados (apéndice B).

PARTE PRIMERA
BASES TEÓRICAS Y METODOLOGÍA

1. LA LINGÜÍSTICA COMPUTACIONAL Y LA LINGÜÍSTICA DE CORPUS

1. Introducción

El papel que desempeñan el lenguaje (facultad innata y sus producciones) y las lenguas (códigos convencionalizados por distintas comunidades) en las actividades humanas, confiere a las genéricamente conocidas como ciencias del lenguaje un carácter marcadamente multidisciplinar. En el seno de este conglomerado de ciencias auxiliares o conexas[3], durante las tres últimas décadas, el procesamiento del lenguaje natural, -o como se prefiere llamarla en el marco de este trabajo, la Lingüística Computacional-, y la Lingüística de Corpus destacan como la consecuencia natural del advenimiento de las tecnologías de la información y de la comunicación.

Como tal disciplina, la Lingüística Computacional ha llegado a alcanzar unos niveles de madurez o de estructuración interna elevados, pese a que todavía puedan producirse dudas sobre su identidad como las que expresa Tordera (2010): "La Lingüística computacional es una disciplina científica relativamente joven. Quizás por ello, esta disciplina presenta problemas tales como su designación (i.e., ¿cómo llamamos a esta disciplina?), su definición (i.e., ¿qué es la Lingüística computacional o qué no es?) o su delimitación (i.e., ¿es una disciplina de la Lingüística, de las Ciencias de la Computación... o de la Filosofía o de las Matemáticas?; y si es una disciplina lingüística, ¿qué lugar le corresponde dentro de la Lingüística?)."

[3] Fabien Morvan, Stéphane Frigo, y Michael Piera incluyen en el campo de las disciplinas tradicionales: la fonética, fonología, morfología, sintaxis, lexicografía, semántica, pragmática, y los estudios sobre el discurso; y entre las de carácter interdisciplinar: la psicolingüística y el estudio cognitivo, la sociolingüística, la etnolingüística, el procesamiento del lenguaje natural y la ingeniería lingüística. En: *L'avenir de la recherche en sciences humaines et sociales: contribution des Sciences du langage*. Leído en <http://www.ecole-doctorale-cli.org/> [Consulta : 12/01/2012].

En cuanto a la Lingüística de Corpus, al igual que la anterior, es en sí misma una generalización, pues en su origen fue una especie de alianza entre corpus, lexicografía y enseñanza. Posteriormente, sin embargo, se convirtió en la herramienta básica de la Lingüística Computacional. Dicha circunstancia dio lugar a que la Lingüística de Corpus ampliara sus objetivos tradicionales, -que se limitaban a validar la coherencia de una teoría elaborada sobre una base introspectiva-, para adaptarlos a los requerimientos de la sociedad de la información.

La colaboración entre las dos disciplinas es muy estrecha, ya que la Lingüística Computacional, por su parte, aporta a la Lingüística de Corpus muchas de las aplicaciones y herramientas que sirven para la explotación y manejo de los corpus como, por ejemplo, aquellas que se ocupan de la anotación y etiquetado semiautomático, del análisis cuantitativo con técnicas estadísticas, o de la búsqueda de concordancias internas.

En lo que concierne a sus objetivos respectivos, hay que destacar que la Lingüística Computacional se ocupa de la creación de modelos computacionales que aplica básicamente a dos campos de actividad: el estudio del lenguaje, con el propósito de mejorar la interacción hombre-máquina, y el desarrollo de herramientas útiles para el procesamiento del lenguaje natural.

En cuanto a la Lingüística de Corpus, se preocupa por crear grandes colecciones de textos con muestras reales de lengua, que son almacenados digitalmente, y con los que se realizan distintos tipos de estudio y análisis empírico de la lengua.

La correlación compleja entre ambos campos de actividad científica explica que pueda existir una confusión sobre la especificidad como ciencias autónomas de la Lingüística Computacional y la Lingüística de Corpus, e incluso que algunos autores las consideren como una única ciencia.

En el presente capítulo, realizaremos una aproximación a lo que, al menos a efectos expositivos, nosotros trataremos como dos disciplinas con entidad propia. Repasaremos sus enfoques respectivos haciendo un especial hincapié en las aplicaciones de la Lingüística Computacional que sirven de ayuda a la Lingüística de Corpus y, finalmente, efectuaremos un breve recorrido por la historia de la Lingüística de Corpus y sus principales aplicaciones.

2. Aplicaciones generales de la Lingüística Computacional

La Lingüística Computacional aparece ya desde los años 60, cuando se crea la *Association for Computational Linguistics*. En 1986, Grishman la definía como la ciencia que se encarga "del estudio de los sistemas de computación utilizados para la comprensión y la generación de las lenguas naturales". Moreno Sandoval, por su parte, se refirió a ella como la ciencia "que trata de la construcción de sistemas informáticos que procesen realmente estructura lingüística y cuyo objetivo sea la simulación parcial de la capacidad lingüística humana" (1998).

El fin último de la Lingüística Computacional es emular la capacidad lingüística humana, es decir crear sistemas informáticos que puedan comprender y reproducir el lenguaje casi como lo haría un humano. No obstante, debido a lo ambicioso del objetivo, no puede decirse que se haya alcanzado todavía de manera plenamente satisfactoria. No es menos cierto que a día de hoy existe un largo recorrido teórico, y que numerosas aplicaciones capaces de realizar tareas con lenguaje natural están ya disponibles en el mercado.

La Lingüística Computacional desempeña un papel muy importante en la sociedad de la información, ya que proporciona herramientas y soporte en ámbitos tan variados como los que se cita a continuación:

a. Comunicación entre humano y máquina

Las técnicas propias del procesamiento del lenguaje natural sirven para mejorar la comunicación entre humano y máquina, ya que permiten realizar un uso más eficiente y completo de aplicaciones informáticas y tecnológicas específicas.

Entre las aplicaciones más conocidas de esta categoría, citaremos los sistemas de recuperación y extracción de la información y las interfaces hombre-máquina. Por recuperación y extracción de información se entiende aquellas aplicaciones que tratan la información almacenada en bases de datos textuales. La recuperación de la información es el procedimiento que trata de encontrar en la base de datos la información más pertinente en relación con una determinada

consulta realizada por un usuario. Se ocupa, por tanto, de filtrar y adecuar las informaciones de una base de datos en función de unos criterios de búsqueda establecidos. Las bases de datos suelen contener gran cantidad de información, por lo que para obtener los resultados de búsqueda más adecuados posibles, se requiere una serie de aplicaciones complementarias que contengan las claves capaces de detectar la pertinencia de una información.

Para lograr sus propósitos de eficiencia, la Lingüística Computacional intenta crear sistemas de búsqueda compleja que se apoyan en teorías semánticas y en ontologías. Un ejemplo ilustrativo serían los prototipos de buscadores semánticos para la Web 2.0. como *LexxeAlpha Search Engine*[4] o *Naveganza*[5] (creado por la empresa española Isoco e implantado en la web del Ayuntamiento de Zaragoza).

En cuanto a la extracción de la información, este tipo de aplicaciones se ocupa de interpretar los datos contenidos en el texto y de crear después un formato adaptado para que pueda ser tratada o recuperada por otras herramientas y tecnologías. Encontramos herramientas de extracción de la información en aplicaciones como *Natural Extractor*[6], creada por Bitext, o en gestores de bases de datos terminográficas como *OntoTerm*[7].

Finalmente, hay que mencionar las interfaces hombre-máquina. Estas aplicaciones tienen como tarea facilitar las interacciones entre el usuario y los ordenadores u otras herramientas informáticas. Su objetivo principal es la comunicación efectiva entre la máquina y el usuario en su lengua natural, bien de forma escrita o a través de la voz, sin necesidad de utilizar lenguajes informáticos o instrucciones complejas. Aquellos sistemas que permiten la comunicación escrita están mucho más avanzados que los de habla, aunque, en los últimos años, se ha producido un interés creciente por este último formato, pues proporciona una forma más sencilla y clara de interacción con el usuario y, por tanto, mucho más adaptada a todo tipo de público. En cuanto a la modalidad de sistemas de interfaz hombre-máquina basados en el habla,

[4] Lexxe Alpha Search Engine se puede consultar en http://www.lexxe.com/.
[5] Naveganza se explica en http://www.isoco.com/naveganza.htm y se aplica en la web del Ayuntamiento de Zaragoza:
http://www.zaragoza.es/ciudad/encasa/buscador_Tramite
[6] http://www.bitext.com/en/whatwedo/consulting/con_extraccion.html
[7] http://tecnolengua.uma.es/ontoterm/index.html

citaremos como un ejemplo muy representativo los servicios de atención a clientes de las compañías telefónicas.

b. Comunicación en distintas lenguas

La Lingüística Computacional ha abierto nuevos caminos para la comunicación entre distintas lenguas desarrollando aplicaciones tan llamativas como las de la Traducción Automática (TA). Hay que destacar que la TA fue uno de los primeros objetivos de la Lingüística Computacional. Es también una de las tecnologías que más ha evolucionado, pese a que quizá, y debido en gran parte a su complejidad, no ha alcanzado el nivel de adecuación o de desarrollo que se pretendía.

La Traducción Automática trata de la creación de sistemas que ayudan a la traducción de una lengua a otra de forma automática y sin la intervención humana. Consiste en tomar oraciones de una lengua y producir las mismas en otra, la llamada *lengua meta*. Dicho proceso de traducción se apoya en complejos análisis gramaticales, léxicos y morfológicos que garantizan la concordancia entre ambas lenguas.

La TA puede aplicarse tanto a producciones escritas como a emisiones orales. Si bien esta última variedad se encuentra mucho menos extendida, ya que requiere un sistema de reconocimiento de habla asociado. Pese a los valiosos avances en sistemas de reconocimiento de habla, todavía no se ha logrado producir una aplicación lo suficientemente eficiente como para lograr la extensión masiva de su uso.

Como ejemplos de sistemas basados en TA existen ciertos programas informáticos como *SYSTRAN*[8], *Trados*[9] o *Translate* (capaz de reproducir las traducciones con voz).

c. Herramientas de ayuda a los lingüistas

Las tareas de análisis de datos que, como es bien sabido, realizan los lingüistas, requieren una inversión en tiempo mucho menor cuando son realizadas de forma semi-automática. Por otra parte, la garantía del nivel de coherencia aumenta considerablemente con la ayuda de estas herramientas, ya que un programa o una máquina realizará el análisis

[8] http://www.systran.es/
[9] http://www.trados.com/en/

siempre de conformidad con los parámetros indicados, sin dejar margen a la subjetividad o a diversas interpretaciones.

Entre las aplicaciones destinadas al trabajo del lingüista, existen herramientas para el análisis textual o de manejo de corpus, que sirven para mostrar análisis de frecuencias de palabras, para la realización de estadísticas de distintos fenómenos de la lengua, de recuento de palabras, así como para la visualización de concordancias de corpus o de alineación sonido-texto de los mismos.

Como ejemplos de las citadas herramientas, podemos citar a *Praat*[10], programa que sirve no sólo para llevar a cabo un análisis fonético y fonológico del contenido de un corpus, sino para la transcripción del mismo. Hay que mencionar asimismo los muchos programas de explotación de corpus y, sobre todo, para la visualización de concordancias, como *Microconcord*[11] o el programa *Contextes*[12], asociado al proyecto C-ORAL-ROM, que sirve como plataforma de visualización y también, para hacer algunos análisis sencillos. Como herramienta más compleja podríamos citar a *WordSmith*[13], ya que, además de ser una plataforma para la visualización de concordancias y el análisis de frecuencias, nos permite realizar otras muchas acciones como hacer listas de palabras clave, análisis cuantitativos, etcétera.

Otras herramientas de ayuda al lingüista bastante extendidas son aquellas relacionadas con la escritura y la corrección de textos. En este caso, podemos hablar de correctores de estilo, ortográficos y sintácticos. Basados en un determinado conocimiento lingüístico (análisis morfológico, reglas sintácticas, gramática de base, etcétera), revisan el texto escrito en busca de concordancias erróneas y sugieren posibles correcciones al usuario. Una aplicación muy corriente de estos correctores aparece en la mayoría de los procesadores de texto que utilizamos en la actualidad, como por ejemplo, *Microsoft Word* o *Pages*.

Los lingüistas también han desarrollado con ayuda de la Lingüística Computacional herramientas para el ámbito de la Lexicografía, que se beneficia así de las condiciones de almacenamiento y consulta de datos que proporcionan los ordenadores hoy en día. En este campo, se

[10] Más información de *Praat* en: http://www.fon.hum.uva.nl/praat/
[11] Posibilidad de descargarlo en: http://www.lexically.net/software/index.htm
[12] Consultable en: http://sites.univ-provence.fr/veronis/logiciels/Contextes/index-fr.html
[13] *WordSmith Tools* se puede consultar en: http://www.lexically.net/wordsmith/

producen, por un lado, diccionarios electrónicos con información y descripción lingüística en varios niveles, y por otro, bases de datos terminológicas. Estas bases de datos suponen una gran ayuda, por ejemplo, para los traductores, ya que contienen términos e información sobre los elementos léxicos de un dominio concreto (medicina, jurisdicción, ingeniería…), y además, pueden tener un carácter plurilingüe. Por otra parte, estas nuevas bases de datos se pueden reutilizar convenientemente en otras aplicaciones como ontologías y en el desarrollo de las herramientas de extracción y de recuperación de la información antes mencionadas.

Finalmente, cabe destacar las nuevas aportaciones que la Lingüística Computacional está realizando en el ámbito de la enseñanza. En la actualidad, encontramos numerosos programas que utilizan técnicas de procesamiento del lenguaje natural, como los programas de enseñanza de lenguas, habitualmente conocidos en español como ELAO o Enseñanza de Lenguas Asistida por Ordenador (CALL en inglés: *Computer Assisted Language Learning*). En este caso, se trata de poner a disposición del aprendiente unas plataformas de enseñanza que aúnen distintos recursos multimedia para el desarrollo de los contenidos previstos para su aprendizaje. Existen además otras aplicaciones derivadas como son los analizadores de producciones escritas de aprendientes, plataformas de ejercicios sintácticos o léxicos o incluso, de creación de ejercicios en línea para actividades y tareas propias de Enseñanza Asistida por Ordenador (EAO).

3. La Lingüística de Corpus: breve historia de su génesis y desarrollo

Un *corpus* es un conjunto de documentos reales (textos, imágenes fijas o móviles…) recopilados y ordenados para un fin específico. La Lingüística de Corpus (en adelante, LC) es la lingüística que se ocupa del procesamiento automático de los corpus específicamente lingüísticos, los cuales son definidos por Rastier (2002) como la "*agrupación estructurada de textos integrales, documentados, eventualmente enriquecidos con etiquetado, y organizados de manera teórico-reflexiva teniendo en cuenta discursos y géneros, y de manera práctica, para servir a una serie de aplicaciones*."

La LC es una metodología propia de la lingüística aplicada[14] que fomenta el análisis de la lengua a partir de los datos almacenados en los corpus, o lo que es lo mismo, de muestras reales de lengua recogidas, tratadas y almacenadas digitalmente en grandes colecciones de textos orales o escritos.

Teóricos como McEnery y otros, (2006) señalan que el término *corpus linguistics* aparece por primera vez a principios de los 80, en lo que se supone el resurgir de la LC. Hasta entonces, si bien no se había considerado una disciplina como tal, sí se la reconocía como una corriente metodológica que se ocupaba del estudio de la lengua a partir de corpus aún no digitalizados, y que basaba sus análisis en muestras reales. De hecho, McEnery & Andrew (1996:2) señalan que prácticamente toda la investigación lingüística antes de la llegada de Chomsky se apoyaba en dicha metodología, puesto que todos los análisis se basaban en la observación del uso de la lengua.

Encontramos ejemplos de estudios con metodologías propias de la LC en muchos autores anteriores a 1950 como Boas, y estructuralistas como Sapir, Newman, Bloomfield o Pike (Cf. McEnery et al., 2006). Los trabajos en LC más habituales de entonces se relacionaban con el estudio de la lengua infantil y la adquisición del lenguaje, en los que los corpus estaban compuestos por las anotaciones diarias de las locuciones del niño a lo largo de un período de tiempo (análisis longitudinales). Otros se interesaban por la lingüística comparativa, como Eaton, estudiando la frecuencia de los significados de palabras en distintas lenguas (alemán, francés, italiano y danés). También se incluían algunos trabajos de sintaxis y de semántica sobre listas de frecuencias semánticas o alguna gramática descriptiva temprana del Inglés como la de Fries (Cf. McEnery & Andrew, 1996: 4).

Una mención particular merece el religioso Roberto Busa, considerado el padre la Lingüística de Corpus y del hipertexto, por su decidida contribución a este enfoque con su proyecto de compilación de obras de Santo Tomás Aquino, de carácter multilingüe, que empezó a

[14] Existe cierta controversia entre los autores sobre si la LC puede considerarse una rama de la lingüística o no. Para algunos como McEnery & Andrew (1996) es una pregunta controvertida, ya que no puede compararse a otras ramas como la semántica o la sociolingüística, porque no explica ciertos aspectos del uso de la lengua. Sin embargo, la LC es para estos autores una metodología que puede usarse en casi todas las áreas de la lingüística para el análisis de los aspectos relativos a cada una de ellas.

producirse a mediados de los años cuarenta, y llegó a alcanzar más de diez millones de palabras.

Sin embargo, a finales de los 50 y principios de los 60, esta disciplina (o *enfoque metodológico* entonces) decae como consecuencia de diversas críticas realizadas por Chomsky y otros autores de la corriente generativa respecto a la naturaleza de sus datos y la validez de sus análisis. Sin entrar en detalles, diremos que la visión del lenguaje y de los estudios que tratan de desentrañar su procesamiento resultaba completamente opuesta en ambos enfoques. Para muchos generativistas, la lengua tiene un carácter infinito cuyas características nunca podrán visualizarse de manera significativa mediante la observación de ningún corpus o conjunto de muestras, pues el análisis resultará inevitablemente insuficiente, parcial, e incluso, falso. Además, para hablar del lenguaje (aludiendo al procesamiento), afirman que lo que procede es estudiar *la competencia*, y no el resultado o *realización* (lo que conocemos en inglés por *performance*).

De alguna forma, se produjo una vuelta al eterno debate existente en lingüística entre los racionalistas, que promulgan un estudio a través de la introspección y una teoría sobre el procesamiento mental del lenguaje, y los empiristas, quienes favorecen, por el contrario, el análisis de datos externos y la observación de muestras reales de lengua para promulgar sus hipótesis.

Por un lado, algunas de las críticas de Chomsky resultaban pertinentes en el momento en el que se realizaron, ya que en los años 60 no existía el desarrollo tecnológico actual que permite almacenar grandes cantidades de datos y que proporcionan a la LC su mayor valor. En aquellos años, exceptuando la obra aislada de Busa, el tamaño de los corpus era muy pequeño, reduciéndose a unas cuantas anotaciones o transcripciones de muestras reales de lengua en papel, por lo que, obviamente, nunca podían conformar muestras representativas de una lengua.

Durante las décadas 60 y 70, por tanto, la LC es relegada, quedando representada sólo por algunos estudios de autores pioneros, quienes creyeron decididamente en su metodología.

Ya en los años 80, la situación cambia radicalmente con el advenimiento de las tecnologías de la información y de la comunicación y, sobre todo, de los ordenadores personales. Estas máquinas mucho

más potentes permitían el almacenamiento de grandes cantidades de datos, que podían ser analizados o tratados rápidamente, lo que facilitaba en gran medida las tareas de análisis (no sólo por el ahorro considerable de tiempo, sino por el aumento del grado de objetividad). Por otra parte, el nacimiento de la Lingüística Computacional fue uno de los hechos que, sin duda, propiciaron el resurgir de la LC, logrando su nueva implantación en el ámbito científico.

En las últimas décadas, la LC ha ido ganando en credibilidad y usuarios hasta llegar al estatus actual en que produce no sólo una abundancia de proyectos, estudios y análisis basados en su metodología, sino que cuenta con una abundancia de canales propios de difusión como congresos, listas de distribución, asociaciones y revistas especializadas.

4. La Lingüística de Corpus: usos y aplicaciones frecuentes[15]

La LC, recordemos, se ocupa de la compilación de grandes colecciones de textos (orales o escritos) para proceder posteriormente a sus análisis y estudio. En algunos casos, el fin de dichos análisis es el conocimiento de la lengua, pero en otros, los resultados se incorporan al trabajo realizado por otras disciplinas, como la Lingüística Computacional o la Ingeniería Lingüística, sirviendo de base para la creación de distintas herramientas y aplicaciones informáticas.

En términos generales, podemos decir que la LC está presente en dos ámbitos principales, a menudo relacionados entre sí:

1. Estudios del lenguaje.
2. Ingeniería Lingüística

[15] Realizaremos una descripción detallada de los usos de los corpus en el capítulo referente a la descripción de los mismos (p. 45).

4.1 Estudios de lenguaje

En estudios del lenguaje, los corpus aparecen como una fuente de datos empíricos, sirviendo para el análisis de la lengua y el desarrollo de herramientas computacionales derivadas del procesamiento del lenguaje natural. También están presentes en la investigación sobre análisis del discurso, gramática, pragmática, semántica, y también en estudios sobre el léxico, una de sus aplicaciones primigenias, donde ayudan a los lexicógrafos a encontrar nuevos usos y ejemplos de palabras y expresiones, y son un soporte indispensable para la redacción de diccionarios y otras obras de referencia.

Además, hay que mencionar su contribución a la enseñanza de las lenguas y la lingüística, no sólo mediante su uso directo, sino también, a través de manuales y libros de texto que se nutren de los ejemplos reales de lengua que suministra.

La LC es necesaria también en lingüística histórica, a la que proporciona textos de distintas épocas, que una vez compilados, pueden ser analizados de forma más sencilla y rápida, gracias a las herramientas de explotación. Es útil asimismo para muchas otras disciplinas como la estilística, la psicolingüística, la sociolingüística, la psicología social, etcétera, a las provee de datos reales con los que comprobar y validar las distintas teorías o hipótesis.

4.2 Ingeniería Lingüística

La LC también mantiene una estrecha relación con la Ingeniería Lingüística, que es la vertiente más tecnológica de la Lingüística Computacional, aquella que trata de construir sistemas para el tratamiento de la lengua. En este ámbito, los corpus participan en proyectos de análisis de las partes de la oración, sirviendo de conocimiento de base para la realización de etiquetadores (*taggers*) y analizadores sintácticos (*parsers*), semánticos y léxicos.

Además, se relacionan con la ingeniería de distintos recursos como léxicos o bases terminológicas digitales, y en diccionarios electrónicos y otras aplicaciones lexicográficas.

Finalmente también aportan su conocimiento para otras aplicaciones informáticas destinadas a la enseñanza asistida por ordenador, y a los programas de traducción automática, que, como hemos comentado anteriormente, fueron el objetivo prioritario de la Lingüística Computacional desde su origen.

5. Conclusiones

La Lingüística Computacional y la Lingüística de Corpus son dos campos íntimamente relacionados entre sí, no sólo porque ambos utilizan el ordenador para el tratamiento de la lengua, sino también porque trabajan conjuntamente tanto en los procesos de análisis, como en el desarrollo de aplicaciones. Resolver si deben considerarse o no como disciplinas autónomas no es una cuestión determinante, pues las categorías son creaciones artificiales para trazar fronteras donde realmente no se necesitan, a no ser para ayudar a explicar la complejidad de conjuntos de investigaciones prototípicas. El lenguaje es un *continuum* y por tanto también lo es su estudio. Lo más importante que hay que retener es que, sean prototipos o disciplinas, la Lingüística Computacional se preocupa principalmente por implementar aplicaciones por medio del procesamiento del lenguaje natural, y la Lingüística de Corpus, por crear grandes colecciones de textos digitalizados y utilizarlos para distintos análisis y estudios sobre la lengua, cuyas conclusiones pueden, a su vez, revertir en el desarrollo y concepción de herramientas por parte de la Lingüística Computacional.

No obstante, hay que reiterar que el gran auge de los corpus se debe en parte al auspicio de la Lingüística Computacional, que supone el marco idóneo para la creación de numerosas aplicaciones y herramientas de análisis de corpus, así como de ayuda para su creación y compilación. Los corpus no tendrían su actual estatus en el panorama científico si la Lingüística Computacional no les hubiera proporcionado, por ejemplo, distintos programas de soporte para la transcripción o etiquetadores, y de explotación de datos o analizadores de toda índole. Son este tipo de aplicaciones las que proporcionan a los corpus su mayor visibilidad e interés, convirtiéndolos en importantes bases de datos sobre la lengua actual y su uso en distintos ámbitos o para diversos fines.

La estrecha colaboración entre la Lingüística Computacional y la Lingüística de Corpus ha generado numerosas aplicaciones muy interesantes para la sociedad tecnológica en la que vivimos; desde las más científicas y reducidas al uso en al ámbito de la investigación, -como las herramientas de explotación de corpus o los propios análisis de la lengua que permiten realizar-, a otras muy útiles para la vida diaria, como los correctores de estilo y ortográficos, los sistemas de traducción automática, de recuperación y extracción de la información y, sobre todo, los interfaces hombre-máquina que garantizan una comunicación más efectiva entre todo tipo de usuarios y las máquinas.

2. CORPUS: definición, tipologías y usos

1. Introducción

El advenimiento de las Tecnologías de la Información y la Comunicación (TIC) ha transformado las maneras de proceder en todos los ámbitos de la actividad social o privada, así como las modalidades de aproximación al conocimiento. Esa revolución no podía dejar de afectar a las ciencias humanas en general, y a las ciencias del lenguaje y la didáctica en particular, campos en los que se enmarca la presente investigación.

A medida en que se producía el desarrollo tecnológico, las ciencias que se ocupaban del lenguaje se han esforzado por integrar esos nuevos potenciales, lo que a su vez ha generado la aparición de nuevas disciplinas. Una de ellas es la Lingüística Computacional. Por otra parte, como el lenguaje (artificial o natural) es también la base de la comunicación tanto entre humanos y ordenadores, como entre ordenadores y ordenadores, en las últimas décadas hemos asistido al desarrollo de numerosas aplicaciones y sistemas informáticos que permiten el estudio, el reconocimiento, la reutilización del lenguaje en diversos campos científicos, así como su análisis en términos cualitativo y cuantitativo.

La Lingüística Computacional, definida por Lavid como "un área interdisciplinaria entre la Lingüística y la Informática que se ocupa de la construcción de sistemas informáticos capaces de procesar el lenguaje humano" (Lavid, 2005: 73), arroja nuevas formas de abordar y entender el estudio científico del lenguaje. Con ella, nacen nuevos métodos de estudio, nuevas formas de validación de hipótesis o de análisis del lenguaje con la ayuda de lenguajes de programación y de programas informáticos, que simplifican y, sobre todo, hacen más rápida la tarea del investigador. Entre estas metodologías de estudio y estas aplicaciones, los corpus ocupan un lugar destacado.

2. Definición de corpus

En términos generales, podemos definir un corpus como un conjunto amplio de documentos (textos, imágenes, sonidos…) almacenados con medios informáticos con una finalidad precisa. Por su parte, McEnery et al (2006:5) definen el corpus lingüístico como una colección informatizada de textos auténticos (lo que incluye las transcripciones de la lengua oral), realizada para ser representativa de un determinado género o de las variaciones de una lengua.

Los corpus, pese a haber estado siempre presentes en los estudios lingüísticos, se han perfeccionado y desarrollado aún más con la aparición de los ordenadores. Eso se explica básicamente por tres razones:

1. la capacidad de almacenamiento de datos ha aumentado considerablemente;
2. el manejo y el acceso a los textos o las palabras resulta mucho más ágil y eficaz;
3. las posibilidades de reutilización de los datos y su uso en aplicaciones que facilitan enormemente la tarea del lingüista por su rapidez y objetividad se han multiplicado.

Asimismo, podemos afirmar que el origen del corpus se remonta a los años 60, siguiendo la estela del *Index Thomisticus* del Padre Roberto Busa, quien creó un índice electrónico lematizado de las obras completas de Santo Tomás Aquino (aunque concluyó formalmente la obra a finales de los 70).

En estos años surgen también los primeros corpus escritos electrónicos, como el *Brown Corpus*[16], realizado en los años 60 en la Universidad de Brown por Nelson Francis y Henry Kucera; y los primeros corpus orales, como el compilado en la Universidad de Edimburgo entre 1963-1965 por Sinclair, que contenía 166.000 palabras en conversaciones informales en inglés. Esta primera fase se caracteriza sobre todo por la intención de aprender a construir y mantener los

[16] Uno de sus manuales se puede consultar en: http://khnt.aksis.uib.no/icame/manuals/brown/

corpus de la mejor manera posible. No se caracterizan por ser corpus de gran tamaño, debido, sin duda, a los escasos medios tecnológicos con los que se contaban.

Ya en los años 70 podemos hablar de una mayor consolidación, empezando a propagarse los corpus por otras lenguas distintas del inglés. Sin duda, la continua mejora de la tecnología impulsa su creación, lo que produce que entre los años 80 y el 2000 se produzca la gran eclosión en el ámbito de los corpus. A partir del año 2000 podemos encontrar numerosas mejoras en la compilación y manejo de corpus, con la creación de numerosas herramientas de edición, análisis y explotación de corpus, lo que no sólo facilita la tarea a los expertos, sino que aporta cada vez una mayor calidad del producto.

3. Requisitos principales de los corpus

Más allá de ciertas características específicas en función del tipo de corpus que se quiera realizar, todos deben poseer, al menos unos principios básicos. Entre otros, podemos hablar de:

- Representatividad

Un corpus pretende ser una muestra de una determinada lengua, por lo que, para conseguir ese fin, ha de ser representativo de dicha lengua o de una de sus variedades. Con ello, queremos decir que las muestras de lengua incluidas en el corpus tienen que ser recopiladas siguiendo un diseño concreto para garantizar la proporción y la diversidad necesaria, de forma que el corpus pueda ser tomado como modelo, y que los estudios o hipótesis que se deriven de él resulten adecuados y coherentes.

- Suficiencia de datos

Esta característica podría relacionarse con la representatividad. Sabemos que un corpus representativo y proporcionado es también el que más cantidad de datos contiene (bien formateados y convenientemente tratados). Si nos planteamos usar el corpus para el análisis de la lengua o para ser utilizado en aplicaciones de ingeniería lingüística, un número

elevado de datos nos dará un uso más aproximado de la lengua, resultará más fiable en cuanto a análisis y las hipótesis serán más fáciles de probar. Pero el desarrollo de corpus es una tarea harto compleja, por lo que es difícil alcanzar esa cifra con el trabajo de un solo investigador o con un único corpus. Por otra parte, el número de datos será suficiente dependiendo de los objetivos que persigamos al crear ese corpus y del tipo de corpus que queramos desarrollar.

Por ejemplo, un corpus de aprendientes o un corpus paralelo tendrán mucho menos volumen que un corpus escrito, -que generalmente es más fácil de recopilar-, debido a la complejidad de las tareas de recogida y transcripción de los datos.

- Calidad de los datos

Por calidad entendemos indicadores como que la muestra sea fidedigna, que esté bien formateada y tratada, que no contenga errores, que se realice conforme a una metodología pensada para ese corpus concreto, etc. Pese a que otro de los requisitos necesarios para un corpus sea un volumen importante de datos –al menos, 300.000 palabras-, a veces es mucho más adecuado aspirar a crear un corpus de calidad. Un corpus más pequeño no es menos representativo al tener un menor volumen. Al contrario, pues al carecer de más datos, nos preocuparemos por garantizar la calidad de lo que está representado. Por tanto, tenemos que ser conscientes de que, en general, la calidad ha de ser mayor cuanto más específico o de menor tamaño sea nuestro corpus.

- Homogeneidad de forma

Este rasgo se refiere a que se ha de preconizar la elaboración homogénea de textos en cuanto a forma se refiere. Todos los textos tienen que ser recopilados y formateados de la misma manera para garantizar su reutilización. Asimismo, esta característica puede relacionarse con la necesidad de poseer una estructura interna que clasifique los datos para una comparación más eficaz de los mismos.

- Tratamiento informático

O lo que es lo mismo, almacenamiento y utilización de los datos mediante la aplicación de un formato electrónico. Dicho formato facilita asimismo la transmisión de los datos y su reutilización en herramientas u otras aplicaciones informáticas que lo necesiten. Todo ello dará también una cierta homogeneidad al conjunto de textos, requisito que acabamos de destacar como conveniente para el desarrollo de corpus.

4. ¿Qué puede aportar la observación de corpus orales?

- Hipótesis e ideas sobre la relación existente entre el léxico y la gramática

Al hilo de lo que muestra Sinclair (1991), y que es retomado por Debaisieux (2009), la lengua no está organizada fundamentalmente partiendo de una serie ilimitada de combinaciones entre gramática y léxico, el llamado *open-choice principle*, sino más bien por una serie de correlaciones entre estructuras semi-preconstruidas, muchas veces dirigidas por el tipo de texto que las contenga, y que es lo que se conoce por el *idiom principle*. Muchas veces, para poder darse cuenta de este fenómeno, es necesario observar la lengua a través de programas de concordancias, que nos dan acceso a gran cantidad de datos de contextos muy distintos y que permiten observar regularidades, que de otra manera, no podrían visualizarse.

- Uso y sentido habitual de las palabras y/o expresiones

Con el uso de los programas de concordancias con corpus, podemos darnos cuenta, no sólo del orden o de las relaciones que se establecen entre determinadas estructuras, sino también del uso y del significado más habitual o mejor dicho, más usual de éstas. Podemos verificar así las hipótesis de nuestra propia intuición o comprobar lo expuesto en los diccionarios o en otros métodos de apoyo al aprendizaje de lenguas. Conocer los

verdaderos usos de una expresión o palabra nos ayuda a tener un nivel de lengua más cercano al de los nativos.

- Conocimiento de la estructura textual

La observación de un corpus permite no sólo centrarnos en el detalle, en un mero aspecto o fenómeno lingüístico, sino observar su relación con su entorno (contexto), lo que creará otra serie de relaciones, dependencias o estrategias, que, de otra manera, no podrían ser observadas. Entre estos aspectos podemos citar la prosodia, la organización del discurso y la secuenciación del texto (enunciados), entre otros.

Los estudios con corpus han demostrado lo alejadas que están las formulaciones de la lengua oral y de la lengua escrita. Si bien se evidencia que parten de un tronco común y de un tipo equivalente de secuenciación, cada una de ellas muestra un uso u organización diferente. Así, puede hablarse de gramaticalidad de lo oral, es decir de la necesidad de adecuarse al contexto para que las interacciones verbales puedan cumplir con la función comunicativa. No estamos hablando de una estructura distinta o un modelo de lengua diferente, sino de una estructura con una organización relativamente diferente a la de la lengua escrita.

- Conocimiento del fenómeno de la variación

Los métodos de enseñanza convencionales confrontan al aprendiente a una lengua estándar, a un modelo de lengua rígido y bastante estructurado, donde queda poco espacio para la riqueza o la variación. Todo ello conduce, generalmente, a un uso de la lengua demasiado academicista, pobre y nada espontánea, consistente en imitar los modelos expuestos. El estudio de corpus, con su presentación de múltiples maneras de uso de una misma forma, con estructuras distintas y variadas, nos hace tomar conciencia de la riqueza de la lengua y de lo que supone una producción efectiva en función del contexto en el que se utilice. El trabajar con corpus nos permite llegar a la construcción de modelos de lengua propios, apoyándonos en la observación lexical, sintáctica y pragmática de los datos que contienen. Todo

ello favorece la creación de hipótesis y la elaboración propia de reglas que conducen, como señala Holec (1990a) a *crear discursos*.

5. Tipos de corpus

Los corpus pueden tener muchas formas o tamaños debido a que generalmente su construcción obedece a distintas finalidades. Aquí, proponemos una clasificación en función de sus características básicas:

- Corpus de referencia y corpus ad-hoc (o monitor corpus).

Los corpus de referencia son aquellos que tienen un tamaño fijo, que normalmente abarcan millones de palabras, y que no suelen admitir ninguna continuación. Son corpus que pretenden mostrar una lengua en su conjunto y, por tanto, convertirse en *referencia* para el estudio de la lengua en cuestión. Es el caso de corpus como *el British National Corpus*[17], el CREA (Corpus de Referencia del Español Actual)[18] o el corpus escrito FRANTEXT[19].

Los corpus *ad-hoc* o *monitor corpus* son los que nacen con un fin determinado, en un ámbito concreto, y que pueden tener un tamaño variable, por lo que pueden ser continuamente ampliados en función de las necesidades de su creación. Un ejemplo de corpus *ad-hoc* sería un corpus creado para una determinada investigación, como por ejemplo, el *Corpus d'Orléans* o ESLO 1 y 2[20], de los que hablaremos más tarde en este capítulo.

[17] Disponible en: http://www.natcorp.ox.ac.uk/

[18] El CREA está a disposición de los usuarios en: http://www.rae.es/rae/gestores/gespub000019.nsf/voTodosporId/B104F9F0D0029604C1257164004032BE?OpenDocument

[19] FRANTEXT será comentado más adelante en este estudio. Se puede consultar, aunque no de forma gratuita, en: http://www.frantext.fr/

[20] El proyecto se describe aquí: http://www.univ-orleans.fr/eslo/spip.php?rubrique1

- Corpus generales y corpus específicos o de especialidad

En referencia a su contenido, los corpus generales son aquellos que intentan reflejar un lenguaje específico, una variedad de lenguaje o una lengua concreta en todos sus contextos de uso. Es el caso del *British National Corpus* o del *American National Corpus*[21].

Por otra parte, los corpus específicos se ocupan de mostrar sólo una parte de la lengua o de una variedad, focalizándose en contextos determinados de uso o en usuarios concretos. Puede ser el caso de corpus como los realizados sobre una determinada variedad dialectal, como los antes mencionados *Corpus d'Orléans* o ESLO o de los corpus de lengua académica, como el *Michigan Corpus of Academic Spoken English*[22].

- Corpus escritos y corpus orales

Los corpus escritos son aquellos que recogen colección de textos escritos en soporte digital. Estos corpus pueden basarse tanto obras literarias como otro tipo de documentos auténticos, en función de las necesidades y las características del corpus. Como grandes corpus escritos encontramos algunos de los antes mencionados, como el FRANTEXT.

Los corpus orales, por su parte, son colecciones de textos orales que son almacenados en soporte digital, y con los que se pueden realizar distintos análisis con la ayuda de herramientas informáticas. Los corpus orales suelen estar compuestos de grabaciones con muestras de lengua, que pueden ser espontáneas o no (si se conoce lo que se va a hablar de antemano), leídas (si se lee un texto escrito previamente o se repiten palabras, muy habitual en corpus destinados al estudio de la fonética o fonología). También pueden contener distintos tipos de grabaciones, pudiendo recoger varias situaciones de habla y

[21] Disponible en: http://www.americannationalcorpus.org/
[22] Fácilmente accesible a través de la web: http://micase.elicorpora.info/

entornos o contextos de comunicación (ámbito privado, familiar, académico, medios de comunicación…) y diferentes tipos de producciones por parte de los locutores, como entrevistas, conversaciones espontáneas de temática libre, tareas pedidas expresamente por el investigador, etc. Como ejemplos de corpus orales podemos hablar de C-ORAL-ROM[23], en cualquiera de las lenguas que lo componen.

Los corpus orales, por su complejidad de elaboración (recogida de datos, transcripción del sonido, permisos de uso y alineación sonido-transcripción entre otros estadios de trabajo), suelen ser menos frecuentes que los escritos, así como muchos menores en tamaño. Pese a todo, lo más reseñable en cuanto a corpus orales es que, si son de tipo espontáneo, suelen ser muy valiosos para el estudio del uso que se da a la lengua en realidad. Al hablar de manera espontánea, el locutor no tiene tiempo de preparar aquello que va a decir, por lo que deja fluir su lengua de la forma más natural. El análisis de este tipo de muestra nos puede dar, por tanto, una valiosa información acerca de los mecanismos de producción del lenguaje, el uso más frecuente de determinados aspectos, permitiendo estudios léxicos, semánticos, pragmáticos o morfosintácticos asociados.

Dentro de los corpus orales podemos hablar también de distintos tipos, puesto que las grabaciones pueden ser sólo de audio o sonido, o bien grabaciones de video-audio. En el caso de las grabaciones de vídeo estaremos aludiendo a un tipo de los llamados corpus multimodales. Un ejemplo de corpus multimodal es el realizado en el proyecto SACODEYL[24], que incluye grabaciones de video con entrevistas a jóvenes en varias lenguas europeas.

Ambos tipos de corpus pueden aplicarse a hablantes nativos de la lengua que se esté tratando de describir o bien a hablantes no nativos[25], cuya lengua materna será distinta a la que

[23] C-ORAL-ROM, dada su importancia para el desarrollo de esta investigación, será comentado de manera más amplia a continuación. Disponible en: http://lablita.dit.unifi.it/coralrom/ .

[24] La web de este proyecto aparece en http://www.um.es/sacodeyl/.

[25] Dentro de los corpus escritos u orales de hablantes no nativos, podemos encontrar los corpus de aprendientes de una lengua concreta, como el corpus al que nos

estén utilizando en el momento de recogida de sus producciones escritas u orales. Como ejemplo de corpus de hablantes nativos podemos situar de nuevo a C-ORAL-ROM. Un ejemplo común de corpus de hablantes no nativos podría ser cualquiera de los corpus de aprendientes que se utilizan en adquisición de segundas lenguas. Es el caso de ICLE (*International Corpus of Learner English*)[26], corpus realizado en la universidad de Lovaina y dirigido por Sylviane Granger, para el estudio del inglés de hablantes no nativos en el entorno académico.

- Corpus monolingües o corpus multilingües

Los corpus monolingües contienen muestras de una única lengua. Es el caso de los grandes corpus de referencia como el *British National Corpus* o el corpus FRANTEXT.

Un corpus multilingüe es, por extensión, aquel que contiene muestras de más de una lengua. Podemos ver un ejemplo de este tipo en el corpus SACODEYL[27]: *European Youth Language* (inglés, rumano, francés, español, italiano, alemán y lituano). Este corpus multimodal nace dentro de un proyecto europeo realizado en 2008 en el que toman parte distintas universidades europeas, coordinadas por la Universidad de Murcia. Su objetivo principal es la creación de materiales auténticos para el aprendizaje de lenguas, basándose en un corpus que tiene como tema principal la vida de los jóvenes. Su metodología es la del estudio de textos auténticos, que son a la vez enriquecidos con toda una serie de ejercicios sobre datos extraídos del corpus. Dicho corpus está compuesto de grabaciones en video de jóvenes de entre 13 y 18 años que responden en su lengua materna a distintas preguntas sobre la familia, la vida cotidiana, las vacaciones, la escuela, los proyectos del futuro o asuntos de actualidad. Este proyecto intenta hacer

---

referimos en el conjunto de esta investigación. Este tipo de corpus, por su importancia en el presente trabajo, será definido posteriormente de manera más detallada.

[26] El corpus ICLE se puede consultar en http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm .

[27] Web oficial del corpus Sacodeyl: http://www.um.es/sacodeyl/

más accesible a los profesores los corpus, ofreciéndoles un recurso donde están etiquetados por temas, gramática, léxico, marcadores del discurso y en función del nivel y los criterios establecidos en el Marco Común de Referencia Europeo para las lenguas (MCER). Todo ello les dota de materiales que no sólo cubren el contexto puramente lingüístico, sino también el cultural, y relacionándolo siempre con ejercicios y actividades pertinentes.

Los corpus multilingües, como hemos visto, son aquellos que se ocupan de más de una lengua y pueden ser, a su vez, de dos tipos: corpus paralelos y corpus comparables.

Los corpus paralelos son aquellos en los que están realizados por versiones diferentes del mismo texto en distintas lenguas, resultando así el original y sus traducciones. Están compuestos de un conjunto de textos, pero en distintas lenguas. Por ello, suelen estar alineados (por párrafos, oraciones, palabras) de manera que pueda observarse claramente la relación existente entre las distintas lenguas que componen el corpus paralelo. Este tipo de textos es muy frecuente en terminología y traducción. Un ejemplo de corpus paralelo podría ser cualquiera de los corpus realizados en la Unión Europea con los documentos oficiales que se manejan en sus sesiones.

Los corpus comparables están compuestos de textos en distintas lenguas que comparten una misma temática, origen o extensión, pero que no son directamente traducciones del mismo texto. Este tipo de corpus suele ser muy usado para el estudio y análisis contrastivo de las lenguas. Un corpus comparable podría ser el *Aarhus Corpus, Danish-French-English Corpus in Contract Law*[28], que contiene distintos textos en las lenguas mencionadas dentro de una temática común.

[28] El Aarhus Corpus puede consultarse en: http://web.bham.ac.uk/forensic/links/acad_links.html

- Corpus de aprendientes

Mención aparte por la importancia que tiene para la presente investigación, hemos de describir los corpus de aprendientes como aquellos que, como su propio nombre indica, se encargan de recoger textos escritos u orales producidos por aprendientes de una determinada lengua. Para recoger los datos o muestras que componen el corpus podemos encontrar, entre otras, distintas maneras, como la producción espontánea, la aplicación de determinadas tareas o las entrevistas directas con el investigador.

Además, los corpus orales de aprendientes, si bien son de gran ayuda para el conjunto de la adquisición de segundas lenguas, no son muy frecuentes puesto que tienen una mayor dificultad en su elaboración que los escritos. No obstante, no sólo suponen un esfuerzo mayor, sino una cantidad de tiempo enorme con respecto a los otros tipos de corpus (derivado de la transcripción de los datos, del etiquetado que se plantee hacer, de sus posteriores análisis, etc.).

Los corpus de aprendientes son principalmente usados en el ámbito de la enseñanza, donde a través del análisis de sus datos, podemos mejorar el método de aprendizaje de la lengua en cuestión, conocer los errores más frecuentes de los aprendientes, ver las necesidades específicas de un grupo o definir nuevos métodos y materiales. Y no sólo eso, sino que el análisis de la lengua del aprendiente puede arrojar luz sobre el proceso de adquisición de una segunda lengua, lo cual incide positivamente el desarrollo de nuevas teorías de aprendizaje o de distintos enfoques más efectivos.

La mayoría de los corpus orales de aprendientes que encontramos se han realizado en el entorno del inglés como Lengua Extranjera (EFL, *English Foreign Language*). En menor medida, podemos encontrar algunos en francés y español, pero sin llegar, ni mucho menos, al nivel de desarrollo que tienen para

el inglés. Actualmente, el corpus ICLE[29] (*International Corpus of Learner English*), proyecto dirigido por la doctora Sylviane Granger en la Universidad de Lovaina (Bélgica), es considerado como uno de los corpus de aprendientes más importantes del ámbito europeo, así como uno de los más estudiados en distintos campos como la adquisición de segundas lenguas, la pedagogía o el Inglés para fines específicos.

En cuanto al francés se refiere, encontramos, ante todo, la iniciativa de la Universidad de Southampton, dirigida por la investigadora Florence Myles, para realizar el FLLOC, *French Learner Language Oral Corpora*[30]. Esta base de datos de corpus, realizada en sucesivas fases desde 1993 a 2008 y a través de distintos proyectos de investigación, se compone de un corpus trasversal de aprendientes de francés cuya lengua materna es el inglés. El conjunto de corpus abarca grabaciones de tareas de producción oral (realizadas en entrevistas directas con el investigador) de aprendientes desde los 9 años, así como distintos niveles y años de estudio de la lengua francesa. Cada uno de los corpus que componen el FLLOC abarca una sección de edad y posee una metodología de recogida de datos distinta (en lo que a tarea de producción oral que el aprendiente ha de llevar a cabo). Entre los corpus más importantes que forman parte de esta base de datos nos encontramos con el *Linguistic Development Project, Salford Corpus, Newcastle Corpus, Progression Project* o *Reading Corpus I* y *II* (de hablantes nativos del francés en universidades británicas).

6. La lingüística de corpus en Francia

Los corpus orales y escritos son una realidad desde los últimos veinte años en la mayoría de los países de Europa y Francia no es una excepción.

Aunque nuestro campo de estudio fundamental es en esta investigación el de los corpus orales, no podemos dejar de citar entre los

[29] http://www.uclouvain.be/en-cecl-icle.html
[30] http://www.flloc.soton.ac.uk/

corpus más importantes de Francia por ser el primero y de más entidad, al corpus escrito FRANTEXT.

FRANTEXT[31] no solo es un corpus, sino que es además una herramienta de consulta de fuentes informatizadas sobre la lengua francesa. Está concebido como un gran corpus de textos literarios franceses y como un programa que permite a través de la Web la consulta y la búsqueda de diferentes elementos. Creado por el *Institut National de la Langue Française*, en su origen, pretendía ser una base de datos con ejemplos de lengua destinados a la consulta de los redactores del *Trésor de la Langue Française*[32]. Esta herramienta contiene alrededor de 4.000 obras (unos 210 millones de casos y cerca de mil autores) pertenecientes a los ámbitos de las ciencias, las artes, la literatura, y la técnica recogidas a lo largo de cinco siglos (XVI-XX) y se puede consultar por medio de una suscripción. La herramienta de consulta permite, además, la búsqueda en distintos niveles de complejidad y sobre aspectos gramaticales. Así, una subparte del corpus, la perteneciente a las obras en prosa de los siglos XIX al XXI, unos 127 millones de ejemplos, incluye la posibilidad de consultar el etiquetado gramatical según las partes del discurso.

Por otra parte, en Francia, las investigaciones sobre corpus orales comenzaron muy pronto, aunque no con la sistematización de su desarrollo ni la metodología que se utiliza hoy en día. Dentro del conjunto de corpus creados en Francia podemos distinguir dos etapas: los corpus históricos y los corpus actuales.

Entre los que llamaremos *corpus históricos*, siguiendo las clasificaciones hechas por Cappeau y Seijido (2005), encontramos, en los albores del siglo, las grabaciones realizadas para el corpus *Archives de la parole* (1912), que se conservan y se intentan digitalizar actualmente en la BNF[33] (*Bibliothèque Nationale de France*). Este corpus surge del impulso de Ferdinand Brunot y de las novedades técnicas de aquella época, como el fonógrafo. Brunot pretende grabar y conservar las manifestaciones de la lengua hablada, buscando sobre el terreno aquellas donde "la palabra

[31] http://www.frantext.fr/
[32] Consultable en: http://atilf.atilf.fr/
[33] http://www.bnf.fr/

tuviera el timbre justo, el ritmo impecable y el acento puro"[34], así como manifestaciones de dialectos y *patois*[35].

A lo largo de los años 50, se empiezan a recolectar grabaciones destinadas a la enseñanza del Francés como Lengua Extranjera (FLE). Estos corpus, sin dejar de ser históricos y de talla bastante reducida, dan lugar al *Français Fondamental* y al *Corpus d'Orleáns* (que pasaría después a ser ampliado en el corpus ESLO), basado en las variaciones dialectales de la zona.

El corpus *Français Fondamental* nació en los años cincuenta con la intención de promover el francés en las colonias que iban a convertirse en independientes. Se realizó entre los años 1951 y 1955 y consta de 275 grabaciones (algo más de 300.000 palabras), a las que se le añadieron una serie de textos escritos provenientes de periódicos. De todo ello sólo se conservan las transcripciones, ya que la parte sonora fue destruida. Pese a todo, supone un punto de inflexión en el desarrollo de los corpus orales en Francia, ya que a partir de este corpus, se empieza a tomar en consideración la posibilidad de realizarlos y de darles un uso distinto.

Así, el *Corpus d'Orléans*[36] o *ESLO* 1 y 2 (*Enquêtes Socio-Linguistiques d' Orléans*) es el corpus más importante y representativo de la lengua oral francesa antes de 1980. Consta de grabaciones realizadas desde 1968 y ampliadas sucesivamente hasta 2008, aunque con fines distintos a los de la etapa anterior. Las primeras grabaciones comprenden unas 200 entrevistas a locutores que vivían en Orléans y más de 300 horas de conversación incluyendo, entre otras, grabaciones con la grabadora oculta, conversaciones telefónicas, reuniones públicas y grabaciones de consultas médico-pedagógicas. Su objetivo primordial era el de crear materiales que sirvieran para recursos de enseñanza de FLE basados en textos auténticos para ser utilizados por sus creadores, profesores universitarios británicos, en el sistema de educación público inglés.

Sin embargo, en la segunda etapa del proyecto, ESLO 1, el objetivo cambia, siendo ahora la conservación y la digitalización de las primeras grabaciones, así como la intención de hacerlas accesibles al público en general a través de la red, y especialmente, para los

---

[34] Cita original: « La parole au timbre juste, au rythme impeccable, à l'accent pur » (Cappeau et Seijido, 2005 : 6)

[35] Podemos definir un *patois* como una variante dialectal regional del francés que se habla en ciertas zonas de Francia.

[36] http://www.univ-orleans.fr/eslo/spip.php?article39

investigadores en Lingüística. A la vez, surge una tercera etapa, ESLO 2, llevada a cabo por el grupo CORAL (*Centre Orléanais de Recherche en Anthropologie et Linguistique*), donde se pretende crear un corpus comparable al que dio origen al proyecto, que constaría de 400 horas más de documentos sonoros. El objetivo final de esta tercera parte sería la comparación de ambos corpus desde una metodología variacionista, que pretende mostrar los cambios y las variaciones sufridas por la lengua a lo largo del tiempo, así como otros aspectos sociolingüísticos.

El corpus de Orléans tuvo además otro aspecto novedoso: supuso el principio de un cambio hacia nuevos métodos de enseñanza de FLE. Su impacto fue tan importante que dio lugar al que sería el primer manual de la historia de FLE basado en corpus: *Les Orléanais ont la parole* (Biggs and Dalwood), aparecido en 1976. Se imponía una nueva manera de enseñar y, sobre todo, unos materiales sonoros de apoyo mucho más reales y sin la apariencia de invención o manipulación que tenían los utilizados hasta ese momento. También suponía un nuevo concepto de ejercicios con los que trabajar y que jamás habían sido utilizados hasta ahora, incluyendo, por ejemplo y como elemento más significativo, la transcripción de grabaciones realizadas en la ciudad de Orléans como forma de contacto con la lengua hablada.

Por todo ello no cabe duda en señalar que este conjunto de corpus históricos constituye el testimonio sobre el francés oral más importante antes de 1980.

A partir de los años 80 se crean nuevos grupos de investigación y comienza la descripción sistemática de la lengua oral por parte de lingüistas de grupos como el GARS (*Groupe Aixois de Recherches en Syntaxe*) de la Universidad de Provence, el proyecto de investigación sobre la Fonética del Francés Contemporáneo (PFC) de la Universidad de Toulouse, entre otros. Sin embargo, no existe ningún proyecto de envergadura nacional. No hay todavía una idea de formar un gran corpus nacional que sea una referencia de la lengua francesa. Los corpus más grandes en esta época constan de dos millones de palabras y no corresponden a ningún estándar internacional, al contrario de lo que ocurre en la actualidad.

Mientras que para la lengua escrita existía ya una gran base de datos como el FRANTEXT, la lengua oral sufría aún un enorme retraso. La situación, pese a la gran tradición del uso de corpus, no ha cambiado mucho en los últimos años y nos encontramos con poca cantidad de

corpus en el conjunto del país, muchos de ellos de pequeño tamaño, muy dispersos y no fácilmente accesibles para el gran público, lo que ha contribuido a su escasa utilización en didáctica de las lenguas.

Para cambiar esta situación, la *Délegation Génerale à la Langue Française et aux Langues de France*[37] (DGLFL) se ha puesto manos a la obra y ha desarrollado un programa, llamado *Corpus de la Parole*, *"*en favor de la conservación, la digitalización, la puesta a disposición, la difusión y la valorización de los corpus orales*"* (*Langues et Cité*, 6: 5).

Este programa tiene entre sus acciones:

- La creación de un grupo de trabajo compuesto por lingüistas (de las universidades y del CNRS), juristas, informáticos y conservadores de bibliotecas y archivos para reflexionar sobre las cuestiones teóricas y metodológicas relativas a la digitalización y la explotación de los corpus orales, lo que ha llevado a la redacción de una guía[38] *(Guide des bonnes pratiques*) sobre aspectos jurídicos, éticos y técnicos publicada por el CNRS.
- La realización de un inventario de los corpus orales disponibles.
- El apoyo a distintos proyectos de investigación para la protección, constitución y explotación de corpus orales junto con las federaciones de laboratorios de investigación en Lingüística del CNRS.
- La digitalización de numerosos archivos lingüísticos sonoros. Se trata de digitalizar los fondos sonoros del francés y de las lenguas que se hablan en Francia, valorizarlos creando un portal web que presente los corpus existentes, e integrarlos en una base de datos que agrupe a todos los corpus de dichas lenguas.

El inventario de corpus orales**,** realizado por Magali Seijido y Paul Cappeau[39] bajo la dirección de la DGLFLF, nace ante la necesidad de

[37] http://www.dglflf.culture.gouv.fr/

[38] BAUDE, O. (cord.) (2006): *Corpus Oraux. Guide de Bonnes Pratiques*. Orléans: Presses Universitaires d'Orléans-CNRS Éditions.

[39] CAPPEAU, Paul et SEIJIDO, Magali (2005) : *Les corpus oraux en français (inventaire 2005, v.1.0)*. Délégation Générale à la Langue Française et aux Langues de France. <http://www.culture.gouv.fr/culture/dglf/recherche/corpus_parole/Presentation_Inventaire.pdf>

disponer de una mejor visibilidad de los corpus que ya existen en Francia y de aquellos que están en curso de realización. Este inventario no sólo los nombra, sino que intenta mostrar otras indicaciones sobre ellos como su nombre, el responsable, el tamaño, el contenido y el estado de los datos (soporte que se ha utilizado, estado de conservación) y datos sobre el método de acceso (acceso libre, parcial, limitado…). También ahonda en qué es lo que se puede consultar (sonido, texto, ambos), en qué proporción (extractos, totalidad del corpus) y en qué condiciones (en el lugar de su realización, consultable en línea, etc.) y la persona o el grupo de investigación al que hay que dirigirse para ello.

Ello supone una gran iniciativa, ya que, al carecer de un corpus nacional de referencia, es la mejor manera de contemplar los recursos sobre la lengua oral que existen en el país y poder así utilizarlos de forma eficaz. Este inventario facilita los contactos y el intercambio entre grupos de investigación, permitiendo descubrir así lo que falta y lo que se podría mejorar, ayudando a establecer futuros proyectos de bases de datos que recojan las necesidades de la lengua actual.

Actualmente, hay también otras políticas encaminadas al desarrollo de los corpus orales y al estudio de la lengua oral. Por una parte, encontramos al Instituto de la Lengua Francesa (ILF: *Institut de la Langue Française*), cuya misión específica es la de favorecer, impulsar y desarrollar la cooperación entre los laboratorios que trabajan en el estudio de la lengua francesa, así como poner a disposición de los investigadores un lugar común de recursos, bien sean grandes corpus del francés, bien herramientas o programas destinados a su explotación y análisis.

Por otra parte, se llevan a cabo 21 proyectos nacionales efectuados en coordinación por varios laboratorios que reúnen a cerca de 150 personas. Desde el año 2000, se ha hecho un esfuerzo particular en la mejora de corpus orales, ya que predomina la idea de que se ha producido un retraso notorio en este ámbito con respecto a otros, así como la creencia de que si una lengua no ofrece un gran corpus disponible en línea, corre el riesgo de ser infravalorada. Entre los proyectos más importantes del ILF destacan:

- El proyecto PFC: *Phonologie du Français Contemporain*[40], que pretende crear la más grande base de datos del francés

[40] La página de consulta oficial del proyecto se encuentra en: http://www.projet-pfc.net/

contemporáneo y una de las mayores del conjunto de lenguas de su entorno, consultable a través de Internet. Entre sus objetivos principales están el proporcionar una mejor imagen del francés oral, mostrando su unidad y su diversidad, basándose en la realidad de su uso y en su diversidad geográfica, social y estilística, favorecer un intercambio entre los conocimientos fonológicos y las herramientas de tratamiento automático del habla, y permitir la creación de mejores materiales pedagógicos para la descripción del francés. Además, su fin último es la conservación de una parte importante del patrimonio lingüístico del mundo francófono. En esta idea de ofrecer una visión global y unitaria de la fonología del francés contemporáneo, se pretende dar cuenta de la diversidad socio-geográfica (previendo cincuenta puntos de recogida de encuestas en los que participen una media de diez locutores), la diversidad de registros, los fenómenos fonológicos más reseñables y la importancia cuantitativa (participando alrededor de 500 locutores, lo que supone entre 800 y 1.000 horas de grabaciones).

- El proyecto CLAPI: *Corpus de la Langue Parlée en Interaction*[41], que se encarga de la recogida, digitalización, identificación y explotación de grabaciones en el seno de los grupos de investigación de Lyon-2 (ICAR) y ENS-LSH, entre otros. Su finalidad es la de asegurar la custodia y la gestión de los corpus más antiguos realizados por el laboratorio a través de una plataforma en XML, así como estimular la producción de nuevos corpus de acuerdo con las exigencias teóricas y tecnológicas que existen en la actualidad. Además, en relación con este proyecto, se está realizando una gramática del francés oral en interacción bajo la supervisión de Catherine Kerbrat-Orecchioni (ICAR Lyon).

CLAPI cuenta con unas 350 horas de conversaciones digitalizadas y con 20 horas de conversación alineadas con su correspondiente transcripción en formato XML y siguiendo la convención ICOR (unas 125.000 palabras). Las grabaciones se realizaron en distintas interacciones espontáneas, que no fueron provocadas ni dirigidas por los investigadores del proyecto, en situaciones sociales variadas (desde una conversación habitual a

[41] http://clapi.univ-lyon2.fr/

actividades específicas relativas a un trabajo) y con hablantes tanto nativos como no nativos.

- El proyecto PRAX, que consiste en el desarrollo de herramientas informáticas dedicadas al tratamiento de los corpus orales. En este caso se trata de la herramienta PRAX: *Plateforme de requêtes et d'annotations de corpus en XML*[42], desarrollada por el LPL de Aix-Marseille (de la que no se encuentra actualmente información).

Podemos encontrar un resumen de los corpus más difundidos en Francia en la siguiente tabla:

CORPUS ORALES DE FRANCIA SEGÚN SU CAMPO DE ESTUDIO	
CAMPOS DE ESTUDIO	**NOMBRE DEL CORPUS**
Sociolingüística	BRANCA-PARIS3, PARIS VII, PARIS-X, STRASBOURG-TABOURET-KELLERT, CRFP
Psicolingüística	CHLOE, GRENOUILLE-KERN, NIMH, SPENCER, WEIL
Psicología cognitiva	CREPCO
Lenguaje infantil	CREDIF, NANCY-CANUT, TOULOUSE-3
Didáctica, enseñanza del francés	BELC, CRELEF, ESLO, FRANÇAIS FONDAMENTAL, GRENOBLE 1 Y 2, PARIS-V, RADIO-FRANCE, STRASBOURG-TABOURET-KELLERT
Fonología, prosodia	C-ORAL-ROM, IMPLANTS COCHLEAIRES, LL, LYON III, MALECOT, NANCY, NIMH, PARIS 3 EA 1483, PERPIGNAN, POI, SPENCER, THESOC, TOULOUSE-VERGELY-PREVOT, TOULOUSE-DUVIGNAU
Morfosintaxis	CAFE, CLER, CORPAIX-2, CRFP-1, DELIC, GARS, GRE, HP, NANCY-DEBAISIEUX, NIMH, PARIS 3 EA 1483, PERPIGNAN, POI, SPENCER, THESOC, TOULOUSE-VERGELY-PREVOT, TOULOUSE-DUVIGNAU
Pragmática	C-ORAL-ROM, OZKAN, TOULOUSE-VERGELY-PREVOT
Interacciones	LYON II-GRIC, LYON II-CLAPI, PARIS 3 EA 1483
Análisis del discurso	CAFÉ, MONTPELLIER, NICE-CHAUVIN, TOULOUSE-4, CRFP
Semántica	TOULOUSE-VERGELY-PREVOT, TOULOUSE-DUVIGNAU

[42] Plataforma de búsqueda y de anotación de corpus en XML.

Fonología	PFC
Tratamiento Automático de Lenguas	OTG
Estudios lingüísticos diversos	ACI, CRFP-2
Otros campos	Interpretación simultánea : LEDERER Subordinación : ALLAIRE Diacronía: THESOC Léxico: TOULOUSE-1 Geografía: ANGERS

Tabla 1: Resumen de corpus existentes en Francia. Fuente: Adaptado de Cappeau et Seijido, 2005: 7-8.

6.1 Corpus C-ORAL-ROM y CRFP (Corpus de Référence du Français Parlé)

Mención aparte, por su trascendencia, merecen los dos únicos intentos de corpus de referencia del francés más completos, que están editados y digitalizados dentro del proyecto europeo C-ORAL-ROM y en el corpus CRFP[43] (*Corpus de Référence du Français Parlé*) dirigido por Blanche-Benveniste en el ya extinguido grupo GARS de la Universidad de Provence[44], y posteriormente por el grupo DELIC. La gran diferencia que hay entre ambos reside en que C-ORAL-ROM puede ser consultado por todos aquellos que lo adquieran en la editorial que se ocupa de su distribución, mientras que el CRFP sólo está disponible en su totalidad y para fines de investigación en el laboratorio DELIC (*Description Linguistique sur Corpus*).

C-ORAL-ROM[45] es una colección de cuatro corpus de lengua oral espontánea de cuatro lenguas romances (francés, italiano, español y portugués). Cada corpus contiene, aproximadamente, 300.000 palabras

[43] Toda la información al respecto de dicho corpus se puede encontrar en: http://sites.univ-provence.fr/delic/corpus/index.html

[44] Hay que destacar que la profesora Claire Blanche-Benveniste también partipaba, junto con el profesor Jean Véronis, de la realización del corpus francés dentro del proyecto europeo C-ORAL-ROM.

[45] http://lablita.dit.unifi.it/coralrom/

transcritas. Todo el conjunto suma 772 textos y representa algo más de 123 horas de conversación.

C-ORAL-ROM está disponible en dos versiones, una destinada a los laboratorios y grupos de investigación que quieran trabajar con él con fines académicos, que consiste en 9 DVD (ni comprimidos ni encriptados) distribuidos por ELRA[46]. Y otra, destinada a las bibliotecas y para uso personal en un formato comprimido y encriptado, junto con un libro que contiene la información más importante de cada uno de los corpus.

Por otra parte, en lo que a sus características técnicas se refiere, podemos destacar que las grabaciones se realizaron en circunstancias naturales, en contextos distintos, lo que hace de C-ORAL-ROM una buena representación de lo que podemos entender por lengua oral espontánea, tanto en el plano de la prosodia como de la sintaxis. Además, los corpus han sido enteramente transcritos según las normas ortográficas estándar del formato CHAT[47] y cuentan con una cabecera donde se da información de los locutores (información sociolingüística, como edad, sexo, profesión, etc.) y de la situación contextual (lugar, número de participantes, tipo de grabación...).

Más concretamente, C-ORAL-ROM en su versión francesa, fue dirigido por Jean Véronis en el seno del grupo de investigación DELIC. El corpus oral francés C-ORAL-ROM consta de 206 archivos con 305 hablantes, que conforman en total unas 30 horas de grabación y unas 296.000 palabras. Como en el resto del corpus, podemos encontrar grabaciones de temática variada y en entornos distintos, abarcando desde el plano familiar y privado al público a través de monólogos, conversaciones y conversaciones telefónicas. Todos los textos están acompañados de sus correspondientes transcripciones, que pueden ser consultadas a través del programa de concordancias *Contextes*[48], realizado por Jean Véronis para este proyecto.

Por otro lado, el *Corpus de Référence du Français Parlé*, CRFP[49], forma también parte del conjunto de corpus DELIC, por ser este el grupo de

[46] European Language Resources Association: http://www.elra.info/
[47]Algunas de sus indicaciones, se pueden consultar en: http://repository.cmu.edu/psychology/181/
[48] El programa se puede consultar en: http://sites.univ-provence.fr/veronis/logiciels/Contextes/index-fr.html
[49] http://sites.univ-provence.fr/delic/corpus/index.html

investigación encargado de su creación. El antiguo grupo DELIC[50] está especializado en la elaboración y la explotación morfosintáctica de corpus orales y escritos, y algunos de sus miembros han realizado corpus de gran importancia como la versión francesa de C-ORAL-ROM, antes descrito.

En 1998, comienza el desarrollo del CRFP, que cuenta con unas 460.000 palabras, a las que se añadirán las contenidas en la segunda fase del proyecto, CRPF-2, que recoge grabaciones provenientes de los medios de comunicación. Su objetivo es dar muestra de la lengua francesa oral actual en las principales ciudades del país, y recoger datos representativos de lo que podría llamarse un francés hablado de uso general y corriente.

Para ello, se compone de 134 grabaciones de distintas situaciones de habla (ámbito privado, profesional y público) y con variados locutores (con diferentes edades, nivel de estudios, procedencia, etcétera), que suman cerca de 36 horas y 50 minutos de habla. Estas grabaciones han sido recogidas en 37 ciudades de provincias y en toda la región de París, con una media de tres grabaciones por ciudad. Como otra de sus características principales destacamos su presentación, ya que aparece bajo la forma de alineación sonido-transcripción en su totalidad.

Pese a todo, ninguno de estos dos importantes corpus, C-ORAL-ROM en su versión francesa y CRFP, se ha utilizado para muchos estudios sobre la lengua debido, en parte, a la dificultad de su consulta. Quizá por ello, y como hemos explicado antes, las autoridades francesas han intentado llevar a cabo una homogeneización de los corpus existentes para dotarlos de un mayor acceso, y facilitar que sean convenientemente utilizados en los estudios lingüísticos actuales.

Por todo ello, podemos decir que los corpus orales no son todavía una realidad habitual en el panorama lingüístico francés, si bien se están garantizando iniciativas para concederles la importancia que merecen y mostrar así todo su potencial para el estudio de las lenguas.

[50] Hay que destacar que desde el año 2008, este grupo se ha fusionado con otros de la Universidad de Provence y ha pasado a denominarse TALEP (Traitement Automatique du Langage Écrit et Parlé).

7. Aplicaciones de los corpus

Los corpus, como sabemos, comenzaron a emplearse principalmente en el área de la lexicografía para la mejora de diccionarios. Sin embargo, el desarrollo de diferentes tecnologías y *software* ha traído consigo una generalización de su uso en ámbitos muy distintos.

En este apartado, trataremos de describir los distintos usos que se realizan de los corpus electrónicos en el marco del análisis del lenguaje, de la enseñanza de las lenguas y de la ingeniería lingüística.

7.1 En estudios del lenguaje

Los corpus resultan de gran importancia como fuentes de datos empíricos. Muchas de nuestras hipótesis sobre la lengua se pueden verificar de forma objetiva en los datos recogidos en los corpus, lo que confiere a la investigación un valor suplementario. Hasta ahora, muchas de las afirmaciones que se hacían sobre las lenguas se basaban en intuiciones lingüísticas de los investigadores, pero no eran objetivamente verificadas.

Dentro del estudio del lenguaje los corpus se utilizan en áreas como el aprendizaje y la enseñanza de lenguas, el análisis del discurso, o la lingüística forense, entre otros.

Los corpus son elementos muy versátiles. Por ejemplo, se pueden reutilizar en investigación sobre el habla, pues proporcionan una gran muestra sobre la lengua que se pretende analizar, que cubre parte de las variables de hablantes necesarias (edad, sexo, estudios, etc.) para garantizar un estudio pormenorizado. A través de su análisis en conjunto podemos lograr, entre otros, generalizaciones sobre la lengua en cuestión, encontrar elementos destacados sobre variación lingüística o encontrar rasgos en función de las variables de hablantes. También pueden realizarse con ellos estudios psicolingüísticos y sociolingüísticos (como, por ejemplo, número de enunciados por tipo de hablante en función de su rango de edad, sexo, nivel de estudios, palabras más frecuentes por tipo de hablante…).

Los corpus también se pueden usar en estudios sobre el léxico, la pragmática y la gramática, de forma que no sólo podemos analizar el léxico y los aspectos gramaticales y pragmáticos más utilizados por un conjunto de hablantes, sino desarrollar aplicaciones en función del uso más frecuente o de las características de una lengua, como en el caso de algunos diccionarios o gramáticas para estudiantes de segundas lenguas. A través del uso de corpus pueden hacerse, a su vez, numerosos estudios cuantitativos y cualitativos sobre el léxico, la gramática y la morfología. De todos ellos, lo más importante es la reutilización de esa información en futuras aplicaciones de ingeniería lingüística, como veremos más adelante.

7.2 En el ámbito de la Enseñanza de las lenguas[51]:

a. Uso directo de corpus en enseñanza

Los corpus pueden usarse en enseñanza de las lenguas de una forma directa, esto es, cogiendo muestras del corpus y realizando actividades sobre ellas en función de lo que queramos trabajar (discurso, aspectos gramaticales, léxico, resumen, búsqueda de información concreta…). También puede realizarse un estudio del corpus por parte de los alumnos, de manera que puedan verificar sus hipótesis sobre la lengua o remarcar ciertos fenómenos o aspectos lingüísticos que sean de interés. Pese a todo, estas segundas prácticas suelen ser bastante limitadas con alumnos de nivel básico, por lo que suelen darse, principalmente, en niveles avanzados y en entornos universitarios. La exploración de corpus por los alumnos que conocemos como *Data Driven Learning* (DDL) aún no está muy extendida en el ámbito de la enseñanza, por motivos de dificultad de las tareas y soportes, así como por su exigencia de dotación tecnológica y de un grado elevado de alfabetización informacional. Sin embargo, no podemos obviar que esta visión de la lengua es muy enriquecedora puesto que expone al alumno a frases y vocabulario que no suelen encontrar en los métodos habituales

[51] Este apartado será convenientemente detallado en el capítulo dedicado a la relación de los corpus y la enseñanza, en la página 55.

de aprendizaje de idiomas y, por supuesto, porque se trata de situaciones comunicativas reales.

b. Uso indirecto de corpus en aprendizaje

Un ejemplo de uso indirecto de los corpus en el aprendizaje de lenguas es utilizarlos como base para la elaboración de métodos y diccionarios. Como hemos mencionado anteriormente, algunos corpus se han utilizado como base para el desarrollo de métodos de enseñanza de lenguas y de gramáticas, como las realizadas por Sinclair para *Cambridge University Press* o el primer método de aprendizaje de Francés Lengua Extranjera basado en corpus, *Les Orléanais ont la Parole* (Biggs and Dalwood, 1976).

Pero no sólo eso, sino que también se han realizado otro tipo de materiales como los diccionarios. Así, encontramos un precursor de esta corriente en el *Collins Cobuild English Language Dictionary*, creado ya en 1987 como el primer diccionario basado en corpus y dirigido especialmente a los aprendientes.

Además de esta utilización, podemos hacer un uso indirecto de corpus en enseñanza de lenguas a través de propuestas a los alumnos con ejercicios basados en corpus, en los que realizar determinadas tareas como búsqueda de colocaciones, de concordancias, clasificación de errores comunes en corpus de aprendientes, comparación de uso entre nativos y no nativos, etc., de manera que con su análisis, vayan adquiriendo la propia lengua meta y el uso más adecuado de ella en función del contexto.

No podemos olvidar otra de las aplicaciones más frecuentes hoy en día: la compilación de corpus de aprendientes, es decir, de colecciones de textos orales o escritos realizados por aprendientes, tanto de una segunda lengua o lengua extranjera, como de lengua para fines específicos (como los negocios, el periodismo, la medicina, etcétera). Este tipo de corpus pone de relieve la variedad de lengua utilizada por el aprendiente, cuyo análisis posterior suele redundar en cambios y mejoras pedagógicas.

c. Uso en aplicaciones de enseñanza asistida por ordenador (EAO) y en programas de aprendizaje de segundas lenguas (ALAO-CALL).

Los corpus sirven también de base para distintas aplicaciones dedicadas a la enseñanza de la lengua a través del ordenador por dos razones principales:

1. Los corpus pueden formar parte de las aplicaciones garantizando infinitos ejemplos de lengua auténtica para los distintos ejercicios, tareas o escenarios que puedan crearse.
2. Del análisis cualitativo y cuantitativo de los corpus podemos extraer hipótesis y conclusiones muy adecuadas para el desarrollo de plataformas y otras aplicaciones. Su estudio pormenorizado revelará algunas de las características que influyen en los procesos de adquisición de lenguas, que pueden ser reutilizadas para la implementación de nuevos programas más centrados en las necesidades y especificidades del usuario y totalmente adecuados a sus procesos de aprendizaje.

7.3 En Ingeniería Lingüística

Como mencionamos anteriormente, muchas de las conclusiones de los estudios cuantitativos y cualitativos sobre corpus se reutilizan en aplicaciones propias de la ingeniería lingüística.

Algunos de los recursos que se sirven de la ayuda de corpus forman parte de sistemas de procesamiento del lenguaje natural como los aplicados en traducción automática (programas informáticos de apoyo a la traducción automática), en recuperación y extracción de información (estudios léxicos, semánticos o gramaticales de aspectos de la lengua para un mayor refinamiento de los resultados de búsqueda) o en interfaces hombre-máquina (sistemas de reconocimiento de voz, principalmente).

Por otra parte, también podemos encontrar otro tipo de programas informáticos como etiquetadores morfológicos, analizadores gramaticales, correctores sintácticos o de estilo, que se basan en corpus y que resultan de gran ayuda para la tarea diaria del lingüista.

Finalmente, como ya hemos indicado anteriormente, los corpus empiezan a tener también una cierta aplicación en el dominio de la enseñanza asistida por ordenador y en programas de aprendizaje de segundas lenguas. Este uso puede ser directo, con muestras del propio corpus que se toman como modelo y sobre las que se realizan ciertas actividades (encontramos un ejemplo en el corpus SACODEYL y en la aplicación C-ORAL-ROM-ELE realizada por el LLI-UAM) o bien de forma indirecta a través de su uso en programas de generación de ejercicios y en plataformas completas de aprendizaje de las distintas competencias (producción escrita, compresión oral o escrita).

8. Ventajas e inconvenientes de los corpus

Una de las principales críticas que se ha hecho a los corpus es su pretensión de documento auténtico, así como su intención de reflejar el uso verdadero de la lengua. Según algunos autores, como Lee[52], no podemos considerar a los corpus como documentos realmente auténticos, ya que consisten en una muestra de lengua programada, tratada y, generalmente, basada en unos fines concretos distintos al de la pura comunicación. Un documento auténtico sería para ellos un texto que se ha producido para responder a unas necesidades evidentes de comunicación en un contexto real.

La creación de un corpus, sin embargo, aísla forzosamente al texto de su contexto de origen, y las muestras de lengua que contiene serían, en consecuencia, lo que Widdowson llama el *decontextualized language*. Aún así, la metodología actual de creación de corpus se preocupa por mostrar de forma sistemática el contexto en el que se han producido los textos y la realidad y características del (de los) locutor(es) que aparecen en ellos, alejando ya esta creencia de descontextualización. De hecho, un corpus

[52] Lee, 1995: 324: *"A text is usually regarded as authentic if it is not written for teaching purposes but for real-life communicative purpose"*

tiene la ventaja de poder mostrar, a través de los programas de concordancias, una multitud de contextos distintos, a los que nunca se podría tener acceso sin el apoyo de corpus de referencia.

A todo lo anterior podemos sumarle otras críticas, como la supuesta falta de espontaneidad o naturalidad del documento y el yugo que impone lo que Labov denomina *la paradoja del observador*[53]. Labov considera que la mera presencia del observador forzará a los sujetos participantes en la conversación o entrevista a controlar lo que dicen, actuando de una manera distinta de si el informador no estuviese presente o si se estuviese realizando una grabación. Esta teoría crea problemas importantes pues nos haría pensar que estamos ante una forma de lengua no real, programada y que no refleja fielmente la lengua en estado puro.

Pese a todo, la mayoría de los autores sí consideran los corpus orales generales como documentos auténticos, especialmente en el contexto del aprendizaje de lenguas. Más controversia se genera entorno a los corpus creados en el entorno académico y, por ende, los corpus de aprendientes. Existe la creencia de que el entorno académico es un contexto no real, entendiéndose real como contexto en el que existen unas necesidades de comunicación claras. Se considera que la comunicación es artificial y distinta de la habitual en otros contextos. Pero no es aceptable asociar mecánicamente situación real con comunicación real, ni situación de enseñanza o entrevista académica con comunicación artificial. La principal dificultad radica a la hora de interpretar si la realidad ha de residir en el texto o en el contexto de producción. La clave se encuentra en la idea de *contexto*, como señala Chambers (2009: 18): "Si l'on restreint le contexte aux conditions de la production originale d'un texte, on condamne bien sûr à l'inauthenticité la quasi-totalité de la lecture des textes littéraires, même par les locuteurs natifs.". Por ello, muchos autores aceptan el contexto de la clase como un espacio válido, puesto que allí se desarrolla de forma auténtica la vida de los alumnos y profesores, y es ahí donde surgen unas necesidades específicas de comunicación. El contexto está perfectamente identificado, puesto que es la realidad en la que se inscriben los aprendientes, y quedan por tanto superadas las críticas de descontextualización asociadas a los corpus de gran tamaño y, sobre todo, a los corpus producidos en un entorno académico.

[53] Traducción propia de "*the observer's paradox*".

Otra de las críticas más frecuentes es la que alude a un posible problema de representatividad y homogeneidad de los corpus. En este caso, podemos responder diciendo que ningún corpus, por muy extenso que sea, es plenamente representativo de la totalidad de los documentos disponibles de una lengua. Un corpus nace con una función determinada, para unas aplicaciones concretas a cuyas necesidades debe responder. Más que evaluar su representatividad en conjunto, debemos de hablar de si el corpus se adecua a las tareas o a las aplicaciones que motivaron su creación. Si es adecuado, entonces se considerará un corpus representativo del dominio y cumplirá las funciones para las que ha sido desarrollado.

Las conclusiones que se deducen del análisis de los corpus, aún siendo totalmente correctas si este se ha realizado con garantías de calidad y representatividad, no pueden ser nunca tomadas como las únicas válidas, puesto que siempre habrá una parte de la lengua que no haya quedado representada. Es imposible abarcar toda la representación de una lengua a través de un corpus o de cualquier tipo de análisis o estudio. Dichas conclusiones deben asumirse con el rango de fiabilidad elevada, como representativas de esa porción de lengua o como hipótesis con una alta probabilidad de verificación posterior. Un corpus no puede ser la panacea ni esperar de él que descubra una verdad absoluta, pero supone sin duda un instrumento valioso para respaldar ciertas hipótesis o intuiciones sobre la lengua.

Además, los corpus de referencia son el medio de recrear, de alguna manera, algunas características del que es reputado como el método ideal de aprendizaje de las lenguas: la inmersión. Trabajar con corpus orales y textuales de referencia de una lengua o de hablantes nativos supone una exposición del estudiante a multitud de situaciones de interacciones entre los nativos de la lengua objeto, sin necesidad de desplazarse al lugar donde se encuentra esa comunidad de hablantes nativos.

Por otra parte, los corpus orales suponen un entorno motivante para el alumno pues maneja modelos de lengua distintos de los ejemplos estandarizados y poco naturales que suelen aparecer en los manuales más usados. Aún así, como bien observa Aston (2001b), los docentes deben tener en cuenta que no sería apropiado generalizar en exceso con ejemplos provenientes de corpus, ya que sin dejar de ser válidos, no son

representativos del uso de la totalidad de los hablantes de la lengua objeto.

Entre los obstáculos más habituales que encuentra el desarrollo de los corpus y de aplicaciones o metodologías asociadas a él es la disponibilidad de los mismos, ya que la mayoría no son accesibles para el público en general, quedándose en la intimidad de laboratorios y siendo exclusiva su consulta para los especialistas con fines académicos o científicos.

También podemos acusar la falta de herramientas gratuitas para el manejo de corpus, tales como etiquetadores, programas de consulta o de búsqueda de concordancias. Programas que, por otra parte, es necesario saber manejar para poder sacar el máximo partido a la consulta que se esté realizando. Con lo que a la necesidad de un cierto conocimiento de la lengua de estudio se le ha de sumar la de un conocimiento informático relativo.

No debemos olvidar finalmente que aunque los corpus pueden resultar estimulantes para los estudiantes por los motivos antes aducidos. En un principio, la tendencia general es de rechazo, tanto porque rompen con las rutinas de la enseñanza tradicional, como por el volumen de nueva información que contienen. Lo mismo ocurre con los profesores, puesto que la mayoría identifica el uso de corpus con una tarea titánica, difícil de manejar para alumnos de cualquier nivel (y especialmente de los niveles más básicos) y muy costosa en cuanto a materiales o sus herramientas de explotación.

Sin embargo, lo que debe hacernos reflexionar es que, pese a todos los obstáculos y críticas que se le oponen, el trabajo con corpus está cada vez más extendido, sobre todo en países que cuentan ya con un largo recorrido de desarrollos y aplicaciones, como es el caso de Francia, Bélgica o Inglaterra.

9. Conclusiones

Un corpus lingüístico es una colección informatizada de textos auténticos realizada para ser representativa de un determinado género o de variaciones de una lengua.

En los últimos años se han desarrollado toda una serie de metodologías y aplicaciones para corpus, ya que los investigadores han sido conscientes de su enorme potencial. Pese a las críticas de escasa representatividad, de falta de contexto y de espontaneidad, los corpus son una fuente de datos auténticos de gran magnitud, que permite una gran variedad de análisis y de estudios de una o varias lenguas. Así, los corpus, al estar almacenados informáticamente, permiten una reutilización exitosa en gran cantidad de aplicaciones de ingeniería lingüística como traductores automáticos, servicios de reconocimiento de voz, programas de interacción hombre-máquina, etc.

Por otra parte, los corpus resultan de gran interés en el ámbito de la enseñanza, en general, y en la adquisición de lenguas extranjeras en particular. Suponen una nueva manera de explorar la lengua que se ha de conocer, aparte de mostrar con claridad su forma de uso más extendida, mediante ejemplos no artificiales, extraídos de la alocución de hablantes nativos o aprendientes reales. Todo ello abre nuevas vías para la enseñanza, con la posibilidad de crear métodos o actividades distintas basadas en corpus, mucho más cercanas a las necesidades de los aprendientes y basados en el uso real de la lengua, consiguiendo mejores resultados. Y sobre todo, se trata de un enfoque que atiende más que los convencionales a la función de comunicación de una lengua.

Pese a los beneficios y las nuevas vías de estudio que los corpus incorporan al panorama de la Lingüística, éstos todavía no están desarrollándose en nuestro país al mismo nivel de otros países europeos. Generalmente, el número de corpus escritos supera al de orales, mucho más difíciles y costosos de realizar debido a la exigencia de recogida de producciones espontáneas, y de la obtención de permisos para la transcripción de dichas grabaciones, su publicación y tratamiento, y también porque la metodología de enseñanza (objeto principal de esta investigación) no está aún suficientemente desarrollada.

3. EL USO DE CORPUS EN LA EDUCACIÓN

1. Introducción

El desconocimiento de la existencia de los corpus y el carácter específico de las competencias tecnológicas que requiere su manejo ha retrasado su incorporación a las prácticas habituales de enseñanza y de la investigación pedagógica, entre las que siguen relegados a un discreto segundo plano. Sin embargo, los repositorios de muestras reales y variadas de una determinada lengua suponen una herramienta muy valiosa para la enseñanza en general, y muy especialmente, para la enseñanza de las lenguas.

En el presente capítulo, hablaremos de la utilización incipiente de los corpus en la educación en general, y en el campo de la adquisición de las lenguas en particular. Nos extenderemos sobre los métodos que han ahondado en su utilización en este segundo campo, así como sobre sus ventajas e inconvenientes más destacados. Posteriormente, nos haremos eco de las investigaciones y aplicaciones más recientes en dicho ámbito, tanto en España como fuera de nuestras fronteras, para finalizar formulando ciertas reflexiones sobre el uso de los corpus orales como base para el aprendizaje del francés como Lengua Extranjera (FLE).

1.1 El uso de los corpus en la investigación pedagógica

La utilización de los corpus proporciona a los investigadores una nueva, estimulante y casi inagotable vía de exploración de las formas más próximas a la lengua real que utilizan los hablantes en cualquier contexto de actividad. Una muestra del impacto social de aplicaciones generadas por la tecnología de corpus lo constituye la evidencia de que muchos usuarios realizan de forma espontánea consultas en corpus, por ejemplo, cuando buscan en Internet información gramatical, ortográfica o semántica de una determinada palabra u expresión. De hecho, hay autores que consideran la Web como un gran corpus, y existen incluso

varias herramientas para crear corpus y manejar concordancias a partir de la Web, como *WebBootCat*[54].

Sin embargo, solo excepcionalmente los corpus han sido utilizados en investigación pedagógica distinta a la del campo de las lenguas. Susan Conrad (1996) es un ejemplo. Esta autora estudia la aplicación de los corpus al campo de la biología. Con el fin de analizar la adecuación de la descripción de las características esenciales de esta disciplina en los textos académicos y científicos, Conrad realizó un estudio del lenguaje utilizado por los libros científicos de biología, comparándolo con los textos académicos que sus alumnos redactaban. Se trataba, principalmente, de dejar en evidencia la gran divergencia que existe entre la lengua académica y la lengua científica de una determinada disciplina. Sin duda, una grave incongruencia, si tenemos en cuenta que mediante el uso de la primera, lo que se pretende es que los alumnos alcancen una buena competencia en la segunda.

Del mismo modo, la utilización de los corpus podría favorecer una mejora de las herramientas de investigación de muchas otras áreas del conocimiento, ayudando en cada caso a la concreción de ejemplos conformes a la realidad objeto de estudio, pero ese potencial de la aplicación de la Lingüística de Corpus está obviamente pendiente de exploración.

1.2 El uso de los corpus en la enseñanza

La Lingüística de Corpus proporciona los materiales necesarios a la generación de herramientas pedagógicas potentes, rápidas, poco costosas y muy accesibles, por lo que las posibilidades de mejora que ofrece a las prácticas de enseñanza de muchas disciplinas es un hecho irrefutable. Sin embargo, su introducción es generalmente muy limitada, y muy poco frecuente fuera del ámbito universitario, donde su uso es propiciado por algunos grupos de investigación ligados a la tecnología de corpus.

Muchas son las causas que podemos identificar para esta escasa aplicación de los corpus en la enseñanza, pero destacaremos principalmente las siguientes:

[54] http://www.fi.muni.cz/~thomas/corpora/IALS/bootcat.htm.

1. EL DESCONOCIMIENTO POR PARTE DE LOS PROFESORES

Fuera de unos círculos reducidos, lo habitual es un total desconocimiento de la existencia de los corpus por parte de los docentes. Muchos de ellos no saben qué es un corpus, para qué sirve ni qué ofrece. Con este panorama, resulta difícil que los corpus se generalicen, pues lógicamente el desconocimiento se acompaña de un completo desinterés por este nuevo enfoque, como así lo constata Boulton (2009b).

2. EL REQUERIMIENTO DE COMPETENCIAS ESPECÍFICAS

A lo largo de su proceso formativo, ni los docentes ni los aprendientes han tenido acceso a las competencias lingüísticas e informáticas específicas mínimas que exige el manejo de corpus. Por tanto, la primera vez que se enfrentan a una actividad que implica el uso de corpus, todos ellos se ven con graves limitaciones para seguirla, puesto que es imposible aprender todas las competencias necesarias en el mismo momento de ponerlas en práctica. Si bien es cierto que autores como Boulton (2008, 2009, 2010), Flowerdew (1998) o Johns (1991) señalan que con unas pocas sesiones de iniciación para mostrar qué es un corpus y el manejo de las herramientas necesarias para su utilización, ya se consiguen buenos resultados.

Desde luego, el aumento progresivo de la utilización del ordenador y de las nuevas tecnologías en las aulas está dotando a los alumnos de unos conocimientos distintos a los de sus antecesores, por lo que, en un futuro no lejano, el uso directo de corpus será más fácilmente asimilables, y podrán manejarse de manera más ágil. Los beneficios de la popularización del uso del ordenador se superponen a la extensión de la formación en competencias tecnológicas generales en los *currícula* de la enseñanza obligatoria en nuestro país. Sería muy deseable, desde luego, que dicha formación se extendiera a los planes de estudio del resto de niveles educativos.

Finalmente, una formación tecnológica a nivel de los estudios de grado de los futuros docentes podría solucionar en parte el citado problema de escasez de competencias. Es de lamentar el arraigo de las metodologías llamadas tradicionales, cómodas para el docente, pero ineficaces para los resultados de aprendizaje, que se centran en el aprendizaje sistemático de la gramática y la repetición mecánica de modelos de lengua. Esos atavismos son naturalmente un obstáculo para la introducción de prácticas innovadoras como los posibles nuevos enfoques relacionados con corpus.

3. INSUFICIENCIA DE RECURSOS MATERIALES Y ECONÓMICOS

Por un lado, numerosos docentes con los que hemos trabajado aducen problemas técnicos del material informático de los centros educativos para desarrollar actividades de este tipo. Muchos se quejan de equipos obsoletos o insuficientes para el número de alumnos. Por otro lado, otros docentes alegan la ausencia de presupuesto para poder comprar las licencias necesarias para la obtención de los corpus de referencia o de los programas de explotación, de concordancias u otro software asociado.

Admitiendo que estás razones son de cierto peso, hay que subrayar que las actividades relacionadas con corpus pueden desarrollarse perfectamente utilizando corpus en línea gratuitos, y que también puede accederse a programas de explotación igualmente libres de licencias y fáciles de encontrar a través de la Web. Son muchos los investigadores que ponen a disposición de los interesados sus programas para manejo de corpus (por ejemplo *Wordsmith Tools*, *Microconcord, UAM-Corpus Tool*[55], etcétera), y también muchos los corpus de referencia actualmente disponibles en línea de forma gratuita. Además, y como ya hemos mencionado, algunos expertos empiezan incluso a considerar la propia Web como un corpus (aunque con algunas restricciones, debido a la información incontrolada que acumula), de modo que han ido apareciendo programas para crear pequeños corpus específicos basados en la red, de muy fácil utilización.

[55] Se puede conseguir en: http://www.wagsoft.com/CorpusTool/

Finalmente, se puede aprovechar los beneficios de los corpus sin necesidad de hacer uso del ordenador en el aula. Como señala Boulton (2010b), se pueden llevar a cabo actividades con corpus seleccionando una serie de concordancias previamente y distribuyendo los resultados del sistema de búsquedas y los ejercicios relacionados en papel. De hecho, Boulton considera que se trata de una opción bastante recomendable para las primeras sesiones de introducir a los alumnos en el manejo de corpus, la metodología y la visualización de resultados (como sabemos, los programas de concordancias nos muestran los resultados de forma distinta a la que estamos acostumbrados, con una concordancia por línea y con la palabra o grupo de palabras objeto de búsqueda en el centro).

4. LA INSUFICIENCIA DE ESTUDIOS CONCLUYENTES

En la bibliografía específica sobre el uso de corpus en el ámbito de la enseñanza se tiene conocimiento de numerosos estudios y proyectos de investigadores de todo el mundo, que intentan valorar el impacto en los aprendientes de este tipo de enfoques didácticos. Por el momento, como señala Boulton (2010b), la mayoría no han sido capaces de obtener datos concluyentes al respecto. Se piensa que el uso de corpus puede resultar beneficioso, y sobre todo, motivador para los aprendientes, pero se desconoce hasta qué punto dicho impacto sobre los resultados de aprendizaje es realmente significativo, ya que los índices de mejora detectados son sólo escasamente superiores a los resultados conseguidos mediante una enseñanza convencional. Tampoco se tienen datos sobre si todos los aprendientes que realizan actividades basadas en corpus movilizan las mismas estrategias de aprendizaje, y sobre si ello redunda en el mismo tipo de resultados. Unas conclusiones tan imprecisas no ayudan a que los docentes perciban como necesaria o positiva la utilización de los corpus, lo que explica que estos sigan relegados, como ya hemos comentado anteriormente, al ámbito universitario y de la investigación.

5. LAS DUDAS SOBRE LA VALIDEZ GENERAL DEL USO DE LOS CORPUS

Existe una corriente que considera que el uso de corpus no es pertinente para aprendientes de cualquier nivel o estilo de aprendizaje. Tradicionalmente, se ha considerado más ventajoso para estudiantes de niveles superiores y con un nivel avanzado de dominio de la lengua de estudio. Algunos estudios que intentaron indagar los beneficios y desventajas de dichas prácticas en aprendientes de otras características, no han resultado concluyentes (Boulton 2009, 2010; y Landure et Boulton, 2010). Una de las razones principales es que la mayoría de estos trabajos se llevaron a cabo en el ámbito universitario. En España, sin embargo, se han realizado algunos estudios a nivel de enseñanza secundaria, tanto para la asignatura de inglés, como para la de lengua y literatura[56], con resultados más interesantes. En ellos, se ha podido demostrar que gracias a los corpus, los aprendientes dominaban mejor ciertos conceptos, y que obtenían puntuaciones más altas en pruebas escritas tras trabajar la sintaxis con ayuda de corpus.

En general, se observa que tras un rechazo inicial, las prácticas con corpus acaban siendo valoradas positivamente por los aprendientes, con las salvedades que trataremos en el siguiente punto.

6. LAS DUDAS SOBRE LA MOTIVACIÓN DE LOS APRENDIENTES

Muchos docentes aducen que el uso de corpus no resulta motivador o atractivo para los estudiantes en un primer momento. Según ellos, los alumnos quieren aprender la lengua, no cómo manejar un corpus ni las herramientas asociadas. A dicho prejuicio se suma la evidencia de que, al igual que los alumnos, hay profesores que prefieren no complicarse la vida en clase, ni probar con métodos que no conocen ni dominan. En otros casos, como señala Boulton (2008, 2009b), no es

[56] Estudios sobre actividades con corpus en la asignatura de inglés con estudiantes de ESO realizados por Victoria López y análisis sobre la puntuación en todos los niveles de la enseñanza secundaria obligatoria y bachillerato en lengua y literatura por Jorge Roselló, presentados en el III Congreso de la Asociación Española de Lingüística de Corpus, Valencia, Abril 2011 (en: CANDEL MORA, M.A. y CARRIÓ PASTOR, M.L (eds) (2011): *Actas del 3er Congreso Internacional de Lingüística de Corpus. Tecnologías de la Información y las Comunicaciones: Presente y Futuro en el Análisis de Corpus*. Abril de 2011, Valencia.)

bien aceptado que los aprendientes lleven la iniciativa en la clase, y que el docente deje de ejercer de maestro de ceremonias. Dicho cambio de roles genera inseguridad, bien porque los profesores piensan que tendrán que hacer frente a tareas que no dominan, o que tendrán que recurrir a la improvisación dejando al descubierto sus debilidades, bien porque tienen la sensación de perder su estatus de profesor.

Para dar un mayor impulso a este nuevo enfoque, sería necesario vencer los prejuicios de que es objeto, divulgando resultados de estudios longitudinales y pormenorizados con aprendientes. Además, su difusión en conferencias y foros de docentes, en este momento de auge del uso de las tecnologías de la información y la comunicación (TIC) en el ámbito de la educación, así como su inclusión en los currícula de estudios universitarios de los futuros docentes, garantizaría sin duda, que los corpus fueran más conocidos e utilizados y que se perdiera parte del miedo inicial a su uso o aplicación.

A nuestro entender, los corpus son un aliado muy útil y práctico principalmente para el estudio de una lengua. No hay duda de que comportan numerosas ventajas, destacando su carácter novedoso como método de análisis, más centrado en el alumno, y que promueve la reflexión sobre la lengua y un conocimiento más profundo sobre su estructura y su uso habitual en situaciones comunicativas reales.

2. El uso de corpus en la enseñanza/aprendizaje de lenguas

A partir de popularización del ordenador, los corpus se convirtieron en un recurso indispensable para muchas de las investigaciones sobre Lingüística y otras disciplinas asociadas como la Metodología de Enseñanza de las Lenguas, la Sociolingüística, o la Lingüística Computacional. Al reunir gran cantidad de datos con medios informáticos, los corpus son un aliado perfecto para el estudio del lenguaje y para el desarrollo de herramientas computacionales de todo tipo que implican el procesamiento del lenguaje natural. Son, sin duda alguna, la mejor fuente de datos empíricos sobre una lengua que se puede encontrar. Dichos datos son la base para establecer hipótesis,

conclusiones y análisis objetivos, fácilmente extrapolables, e independientes de la perspectiva e intuición lingüística personal del investigador que tanto determinó en su momento las primeras investigaciones en Lingüística.

Como ya queda apuntado, los corpus se utilizan en investigaciones sobre el habla, en estudios lexicográficos, en análisis de la gramática y del discurso, siempre ayudando a la cuantificación de la variación y como medio para validar las hipótesis de las distintas teorías lingüísticas. En todos los casos, los corpus han tenido una contribución muy significativa, introduciendo cuestiones esenciales como la cuantificación y la explicación de estructuras o de patrones sintácticos.

Era inevitable, por tanto, que la enseñanza acabara aprovechando el potencial de aplicación de los corpus y que, poco a poco, se fuera revelando su utilidad para diferentes formas de consulta y estudio (de manera directa o indirecta, como observaremos más adelante). La mayoría de los recursos pedagógicos son la réplica del diferente peso de lo racional o lo empírico que emana de las distintas teorías lingüísticas y enfoques pedagógicos. Si observamos los libros de texto, comprobaremos que la mayoría contiene ejemplos o textos inventados para la ocasión que, habitualmente, están muy alejados del uso real de la lengua. Sin embargo, en otros casos menos frecuentes, los modelos de lengua que se presentan son más cercanos a la realidad, pues están basados en estudios cuantitativos y de frecuencias de uso o, simplemente, se trata de textos y audios auténticos procedentes de distintos corpus, o de fuentes reales como periódicos, televisión, cine, entrevistas… Es a través de este tipo de materiales que los corpus encontraron su vía de inserción en la enseñanza. Todo ello se vio favorecido por la expansión de enfoques basados en la utilización de las TIC.

En la enseñanza-aprendizaje de lenguas, los corpus han sido utilizados especialmente para ofrecer ejemplos ilustrativos de un determinado conocimiento extraídos de producciones auténticas. Este tipo de uso acerca las clases de lengua a situaciones comunicativas de la vida real, y a la utilización habitual de la lengua por parte de los hablantes nativos.

El incremento de los corpus en el ámbito de la Lingüística Computacional, unido a la efervescencia de nuevas teorías y enfoques de enseñanza-aprendizaje, favoreció efectivamente la apertura paulatina de

nuevas vías de aplicación (estudios cuantitativos o cualitativos), siendo la enseñanza de lenguas uno de esos campos pioneros en los que los corpus son utilizados tanto como base para el análisis y la comprobación de hipótesis sobre los procesos de aprendizaje, como para las prácticas de lenguas de estudio.

Los corpus, o colección de textos escritos y/u orales con abundantes ejemplos de lengua, son para el aula de lengua un documento auténtico más. Sus distintas funciones y variados niveles pueden ser utilizados eficazmente, siempre que se tenga muy en cuenta el objetivo de su empleo, y que se proporcione a los alumnos los conocimientos necesarios para su buen entendimiento y aprovechamiento.

La Metodología de Enseñanza de las Lenguas es una disciplina nacida, como es bien sabido, hacia los años 50, que ha experimentado un importante desarrollo durante las últimas décadas gracias a los sucesivos avances tecnológicos y metodológicos. Una mayor actividad científica en el citado campo ha ido generando la aparición de numerosas teorías e hipótesis sobre los mecanismos de adquisición para todo tipo de aprendientes, sobre la lengua característica del aprendiente o interlengua, y sobre los enfoques más favorables al desarrollo del aprendizaje. Todo ello unido al aumento de la difusión de tales ideas en foros especializados ha tenido un impacto significativo en la evolución de las prácticas docentes.

Es evidente que los aprendientes necesitan interactuar en la lengua meta para poder adquirirla correctamente (Larsen-Freeman and Long, 1991; Chapelle, 1997). Los ordenadores y, por extensión, las aplicaciones que requieren su uso como los corpus, pueden ofrecer a los aprendientes nuevas y excelentes oportunidades de interacción, permitiéndoles que observen cómo se usa la lengua en un ambiente natural. Los corpus, por la gran cantidad de datos que contienen, sus posibilidades de actualización y desarrollo continuo, y la facilidad de su consulta por medios informáticos, son un soporte idóneo para las prácticas de enseñanza.

La MEL, como decimos, es un lugar de encuentro de diversas disciplinas y enfoques que dialogan entre sí en busca de una mayor eficacia de los procesos de enseñanza/aprendizaje de las lenguas. El uso de corpus ha sido una de esas aportaciones que se produjo entre los ochenta y noventa, con el nombre de TALC (*Teaching And Language*

Corpora)[57]. El TALC se generó como consecuencia de encuentros entre expertos que comenzaron a debatir sobre los posibles beneficios del uso de corpus en entornos educativos. La primera *TALC Conference* bianual tuvo lugar en Lancaster en 1994, y su décima edición se desarrollará en Julio 2012 en Warsaw (Polonia). La utilización de corpus en la enseñanza/aprendizaje de lenguas no es por tanto un hecho reciente[58]. Pese a ello, es preciso constatar que, a pesar de que muchos autores preconizan su uso para la enseñanza de segundas lenguas, pocos son los profesores que le dan un uso real.

El grupo de investigadores agrupado en torno a TALC reflexiona sobre todo lo necesario para la realización de corpus, así como sobre el lugar que estos deberían ocupar en las estrategias de enseñanza de las lenguas. En los últimos veinte años, la utilización de corpus se ha extendido a muchas investigaciones universitarias, aunque no todas las lenguas se benefician por igual. Destacan sobre todas, las investigaciones con corpus para el inglés, como lengua materna y como lengua extranjera. Los citados estudios cuentan no sólo con investigaciones de prestigio, sino que existe toda una red de conferencias, publicaciones y asociaciones dedicadas en exclusiva a la descripción de corpus en lengua inglesa. Lo cierto es que la disciplina TALC no sería lo que es sin ese importante volumen de estudios sobre el inglés. En muchas ocasiones, hablar de TALC es hablar casi exclusivamente de inglés como lengua extranjera (EFL, *English as Foreign Language*). De hecho, la inexistencia de conferencias y de revistas especializadas, e incluso de terminología específica para referirse a la investigación sobre corpus, en otros idiomas, es una invitación a reflexionar para la comunidad científica de esas otras lenguas, sobre todo para aquellas que ni siquiera han empezado a trabajar con ese enfoque. No es este el caso del francés y del español. En dichas lenguas, el interés por los corpus aumenta progresivamente, pero sus estudiosos se encuentran aún demasiado aislados y, por lo tanto, sus conclusiones son menos sólidas y están menos homogeneizadas que en el caso del inglés.

---

[57] Antes de considerarse a TALC un nuevo campo de estudio en los años noventa, encontramos algunos exponentes aislados, como el de Peter Roe, en la Aston University en 1969 que usaba corpus para su enseñanza en Inglés para fines específicos (McEnery and Wilson, 1997:6).

[58] El primer método que introdujo los corpus en enseñanza de lenguas en Francia fue *Les Orléanais ont la parole*, que vio la luz en 1976.

En cuanto a la utilización de los corpus en la enseñanza en Francia y España, por ser este el ámbito de aplicación de nuestro estudio, como demostraremos en este mismo capítulo, en Francia se han realizado más proyectos con corpus que en España, y es que nuestro país sigue ignorando este tipo de metodología fuera del ámbito científico y universitario. Destacaremos, no obstante, algunos proyectos que consideramos de gran interés realizados para el español como lengua extranjera y para el inglés.

Sea como sea, es innegable que los corpus son una fuente inagotable de datos de interés sobre la lengua objeto, y son múltiples los servicios que pueden prestar durante el proceso de aprendizaje. Aston realiza la siguiente propuesta de posibles funciones (Cf. Aston, 2001:4):

- ayudar a los aprendientes a validar hipótesis y descubrir características de la lengua y de la cultura meta;
- consolidar lo que ya conocen de ella;
- entender y producir textos de diferentes tipos y en distintos contextos;
- crear y mejorar las habilidades de expresión y comprensión (oral y escrita) necesarias para una correcta interacción en la lengua meta;
- proporcionar nuevos escenarios de interacciones comunicativas para la práctica de la lengua meta;
- crear y modelar la conciencia lingüística de los aprendientes y su autonomía de aprendizaje.

2.1 Diferentes enfoques de aplicación de corpus a la enseñanza/aprendizaje de lenguas

Como ya hemos reiterado, los corpus ofrecen una cantidad prácticamente infinita de datos sobre la lengua de estudio. Por tanto, a través de búsquedas concretas de elementos lingüísticos, o de una exploración al azar guiados por la intuición o las necesidades de conocimiento, tanto los docentes como los aprendientes pueden generar

sus propias entradas, así como integrar datos procedentes de otras fuentes y completarlas o validarlas empíricamente. Por tanto, uno de los primeros usos que suele darse en utilización de corpus para la enseñanza de lenguas es la creación y validación de descripciones lingüísticas. Dicha descripción puede referirse a las colocaciones, aspectos pragmáticos o de estructuras que detentan una función diferente en función del contexto en el que se insertan, estructuras que son propias de un determinado contexto y que lo caracterizan, etcétera.

De hecho, para la enseñanza de lenguas, los corpus pueden utilizarse de múltiples maneras, pero dos enfoques metodológicos destacan como los más habituales: el uso indirecto y el uso directo. Sin embargo, otros autores (Frankenberg-García, Flowerdew and Aston, 2011; McEnery and Xiao, 2010), siguiendo las propuestas de Leech (1997), señalan que existe un tercer uso de corpus: la compilación con fines pedagógicos.

2.1.1 El uso indirecto (o *enfoque Cobuild*)

Un uso indirecto no trabaja con partes del propio corpus, sino que se nutre de los análisis, hipótesis y conclusiones extraídas de su estudio. Así, un corpus se emplea de forma indirecta cuando contribuye a elaborar, mejorar o complementar:

- métodos de enseñanza,
- diccionarios,
- gramáticas y otros libros de consulta,
- programas informáticos para aprendizaje de lenguas,
- Test o evaluaciones de conocimientos.

Algunos autores, como Aston (1997), lo denominan el **enfoque COBUILD**, por ser esta una de las primeras editoriales interesadas en la realización de diccionarios con la ayuda de datos procedentes de corpus.

Como decimos, los corpus aportan una información muy rica y variada sobre una lengua. No sólo nos permiten cuantificar fenómenos, comprobar hipótesis de diferentes teorías lingüísticas o elaborar un

léxico específico para un determinado campo de estudio, sino que inciden en aspectos tan importantes como la variación (no recogida habitualmente en diccionarios y otras herramientas clásicas no digitales) y sobre todo, en el uso real por los hablantes nativos.

Aunque resulta cada vez más frecuente, los corpus no se han usado habitualmente bajo esta modalidad. Generalmente su uso implica un proceso laborioso, y no todos los investigadores lo consideran necesario para un buen conocimiento de la lengua, optando por metodologías más clásicas.

Ya hemos comentado que los corpus pueden contribuir a mejorar los materiales y a crear escenarios comunicativos próximos al entorno nativo. Existen numerosos estudios con análisis del vocabulario de libros de texto y su nivel de adecuación a los niveles de los aprendientes. Biber y Conrad (2001), por ejemplo, realizaron un simple análisis cuantitativo de un corpus de referencia, del que obtuvieron los verbos más frecuentes del inglés. Doce de estos verbos aparecían en un 45 por ciento de las concordancias del registro conversacional, y sólo un 11 por ciento en la prosa académica. Otro análisis de estos doce verbos más frecuentes en los libros de texto, arrojó que la mayoría de ellos no aparecía en las primeras lecciones, ni tan siquiera en materiales de los niveles más básicos. Por tanto, se constata que los corpus son muy útiles para la mejora de los materiales existentes, pues ponen a disposición de los aprendientes un conocimiento de lengua real, que puede servirles para cualquier contexto en que tengan que utilizarla, desde los primeros momentos de aprendizaje de la lengua.

Como ya hemos señalado en el apartado sobre los tipos de corpus, algunos se han utilizado como base para el desarrollo de métodos de enseñanza de lenguas y de gramáticas, como los creados por Sinclair para *Cambridge University Press* o el primer método de aprendizaje de Francés Lengua Extranjera basado en corpus, *Les Orléanais ont la Parole* (Biggs and Dalwood, 1976).

Se han realizado asimismo otro tipo de materiales como diccionarios, tanto específicos para aprendientes, como generales o guías sobre las colocaciones frecuentes. Un precursor de esta corriente es el *Collins Cobuild English Language Dictionary*, creado ya en 1987 como el primer diccionario basado en corpus (utilizando el corpus COBUILD, de 20 millones de palabras) y dirigido especialmente a aprendientes, pues señala la frecuencia de uso. Dicha frecuencia resulta de especial ayuda

para ordenar los principales significados de una palabra, para identificar las colocaciones más comunes y los patrones léxico-gramáticos más utilizados. En los últimos años también se ha publicado el *Longman Dictionary of Contemporary English* (2005), coordinado por Brian Abbs e Ingrid Freebairn y algunos de los primeros diccionarios específicos para estudiantes, como el *Oxford Advanced Learner's Dictionary* (2005), editado por A. S. Hornby, el *Oxford Collocations Dictionary for Students of English*, realizado por Diana Lea en 2002 y el renombrado *Collins COBUILD English Dictionary for Advanced Learners* (2001), implementado bajo la dirección de J. Sinclair.

En cuanto a gramáticas, la primera surgió en 1990 dentro del proyecto COBUILD, y se llamó *The Collins COBUILD English Grammar*. Se trata de una gramática dirigida a los aprendientes donde las categorías gramaticales son fácilmente accesibles y están orientadas desde la semántica. Por tanto, se ofrece al usuario una visión más clara, porque pueden aprender estructuras sintácticas junto con el léxico asociado considerado más frecuente.

Durante los 90, una gran parte de las investigaciones realizadas se centraba en el desarrollo de estas gramáticas, y así aparecieron la *Logman Grammar of Spoken and Written English* (1999), mucho más desarrollada que la de COBUILD, y realizada por D. Biber, S. Johansson, G. Leech, S. Conrad y E. Finegan. Se trata de una gramática de referencia muy útil porque contiene gran cantidad de información cuantitativa y una distribución en función de las estructuras más frecuentes para cada registro de uso del inglés actual. Y, posteriormente, aparece la definitiva *Logman Student Grammar of Spoken and Written English* (2005), coordinada por grandes investigadores de este campo como son Douglas Biber, Geoffrey Leech y Susan Conrad.

Otro ámbito, algo menos extendido, en el que los corpus se usan también de forma indirecta es en la evaluación o tests de conocimientos. El análisis de corpus ha redundado en el desarrollo de tests más adaptados y acordes con las fases de aprendizaje del aprendiente. Por otra parte, muchos de estos tests sirven a su vez de alimentación para los corpus ya existentes, como medio para conseguir nuevas muestras. Por ejemplo, el *Cambridge Learner Corpus* el *Corpus of Young Learners English Speaking Tests* o el *Cambridge Corpus of Spoken English* se nutren de los tests de la *University of Cambridge Local Examinations Syndicate* (UCLES).

Sin embargo, la mayoría de los corpus no se utilizan de forma aislada, sino que forman parte de sistemas o herramientas informáticas más complejas. Hablamos aquí de las plataformas de aprendizaje de lenguas asistido por ordenador, donde sirven para evaluar inicialmente al alumno y como comienzo de su autoaprendizaje, para, más tarde, poder controlar los nuevos conocimientos que va adquiriendo, realizar la corrección de errores y la retroalimentación necesaria para reforzar el aprendizaje. Uno de estos ejemplos lo encontramos en la plataforma *Test-Builder*, de Kaszubi y Wojnowska, para la creación de ejercicios de inglés como L2[59].

Finalmente, menos extendida es su aplicación en libros de texto o manuales para el estudio de la lengua. En parte debido a que se trata de un material costoso de producir, que no tiene un amplio espectro de venta, por lo que no resulta rentable para las editoriales. Muchos de los manuales actuales señalan que se basan en corpus, sin embargo no lo reflejan claramente (Boulton, 2009b), resultando dicha contribución prácticamente invisible. Existen algunos ejemplos de libros que utilizan los corpus, pero como para el resto de materiales de apoyo y referencia, casi todos son para el inglés. Así, podemos citar el *Touchstone* de McCarthy, McCarten y Sandiford (2006), basado en el *Cambridge International Corpus*, pero diseñado casi exclusivamente para el mercado norteamericano, o libros muy específicos como *Exploring Academic English* de Thurstun y Candlin (1997), articulado a raíz de la lista *University Word List* de Nation, y que se basa en el inglés académico de trabajos escritos a nivel universitario.

Posteriormente, los corpus han formado parte de herramientas y aplicaciones de aprendizaje asistido por ordenador (CALL), sobre todo en plataformas de aprendizaje y programas para la creación y realización de ejercicios o tests como *eXXelant* o *MIRTO*, como veremos más adelante.

[59] Disponible en: http://ifa.amu.edu.pl/~kprzemek/DOC/index.html

2.2 El uso directo de corpus en la enseñanza (Enfoque data-driven learning)

Utilizar un corpus de forma directa requiere manejar el corpus a través de las herramientas informáticas y programas necesarios con el fin de realizar investigaciones sobre la lengua. Se trata de un enfoque donde el o los corpus son recursos para el aprendiente, que los maneja, de forma guiada o no por el profesor, a través de programas de concordancia informáticos o con una impresión de ciertas concordancias. A partir de los corpus, puede realizarse distintos análisis sobre aspectos discursivos, léxicos, sintácticos, etcétera.

2.2.1 El Data Driven Learning o DDL

La metodología de trabajo directo con corpus en el ámbito de la enseñanza se conoce en la bibliografía de la Lingüística de Corpus como *Data Driven Learning* (DDL). En español es habitualmente traducida como 'aprendizaje dirigido por los datos', pero existe una gran tendencia a mantener el nombre anglosajón o citarlo con sus siglas, quizá porque desde su origen, la mayoría de los trabajos realizados en este campo se deben a autores anglosajones.

Uno de los investigadores que inició esta metodología es Tim Johns, que realizó sus primeros trabajos en los años 80 y 90 en la Universidad de Birmingham. De él proviene la definición y denominación de este enfoque, que se caracteriza, en sus propias palabras, por:

What distinguishes the data-driven learning approach is the attempt to cut out the middleman … and give direct access to the data so that the learner can take part in building up his or her own profiles of meaning and uses (Johns, 1991:30).

En sus investigaciones, Johns desarrolló programas para extraer concordancias y elaborar estrategias para la clase. Él estaba convencido de que partir de la lengua real, favorecía el aprendizaje. El método de

Johns servía para describir los fenómenos lingüísticos que posteriormente eran analizados por los aprendientes con el fin de inferir y validar generalizaciones descriptivas sobre su uso. Los programas con los que contaba el aprendiente para el manejo de la lengua eran aún muy reducidos. En ellos, el aprendiente seguía teniendo un papel bastante pasivo, ya que se limitaba a formar parte de un sistema de preguntas y respuestas: la Inteligencia Artificial no había descubierto aún cómo crear programas más pedagógicos, basados en las fases de adquisición de una lengua.

Una de las grandes críticas que se hacen a los corpus es el problema de la contextualización. Sin embargo, como ya indicamos en el primer capítulo, el hecho de utilizarlos en un entorno educativo ya es en sí una contextualización y una autentificación. Freda Mishan (2004:222) señala al respecto que una de las formas de autentificar un corpus para el aprendizaje de lenguas es, precisamente, utilizar la metodología del DDL, porque implica a los alumnos en el corpus a través de una serie de tareas concretas.

Por su parte, Johns (1991) afirma que el aprendiente de una lengua es también esencialmente un investigador, cuyo aprendizaje necesita estar dirigido. De ahí el nombre escogido para su enfoque: *Data Driven Learning*. La tarea del aprendiente consiste, básicamente, en descubrir la lengua, y el profesor se compromete a proporcionar un contexto donde pueda desarrollar sus estrategias de descubrimiento e investigación y de "aprender a aprender".

La forma de trabajar con un corpus puede ser también diferente en función de lo que queramos fomentar en el aprendiente:

- dejando a los alumnos trabajar por sí solos, con sus propias intuiciones y verificando sus propias hipótesis;
- a través de una serie de actividades seleccionadas por el guía-profesor para destacar patrones o estructuras lingüísticas concretas.

2.2.2 El aprendizaje autónomo

En el caso del aprendizaje autónomo, centramos la enseñanza en los alumnos, que se convierten en protagonistas de su propio

aprendizaje. Nuestros aprendientes se asemejarían a investigadores, realizando un trabajo de observación directa del corpus, de manera que pueden verificar sus hipótesis sobre la lengua o remarcar ciertos fenómenos o aspectos lingüísticos que sean de su interés. Es decir, realizando su propia investigación. Los procedimientos para el análisis de las muestras de corpus pueden seguir dos metodologías distintas: **inductiva** e **deductiva**.

- La **metodología inductiva** propone que los aprendientes infieran sus conclusiones con los datos obtenidos en los sistemas de concordancias, descubriendo hechos de la lengua a través de la identificación de patrones lingüísticos que validen sus hipótesis previas. Estas hipótesis irán modificándose progresivamente siguiendo procesos de observación, clasificación y validación, de tal manera que se conviertan al final en generalizaciones o conclusiones propias sobre el uso de la lengua. Podemos relacionar esta metodología con la idea de Johns del aprendiente-investigador o de aprendizaje por descubrimiento (*discovery learning*), también mencionado por Bernardini (Cf. Aston, 2001). Los expertos han concedido a esta metodología una gran efectividad porque estimula niveles más profundos de aprendizaje y porque, al implicar de forma directa al aprendiente en el proceso de formación de conocimiento, el aprendizaje se mantiene por más tiempo en la memoria.

- La **metodología deductiva** es aquella en la que los aprendientes aplican las generalizaciones que ya han adquirido para clasificar los datos que proporcionan los corpus, validando así las reglas que conocen. En este caso intentarán consolidar, redefinir, discutir o invalidar las reglas a través de las muestras de los corpus, adquiriendo un conocimiento mucho más adecuado al uso real de la lengua en distintas situaciones comunicativas. Este tipo de actividades suele implicar búsquedas, por ejemplo, de una expresión concreta y de sus usos o de comparación entre estructuras similares para una mejor distinción de los contextos en los que se aplica. En ocasiones, estas reglas no se pueden probar, o necesitan matizaciones, por lo que se produce también un trabajo inductivo al plantear hipótesis adicionales o nuevas reglas de uso.

Además de esta metodología de trabajo, otro uso directo de corpus en enseñanza de lenguas se puede realizar a través de ejercicios en los que los alumnos llevan a cabo tareas como búsqueda de colocaciones, de concordancias, clasificación de errores comunes en corpus de aprendientes, comparación de uso entre hablantes nativos y no nativos, etcétera, de forma que vayan adquiriendo la propia lengua meta en función del contexto. Este uso es también a veces considerado como un uso indirecto si los alumnos no trabajan físicamente con el corpus, sino que ven parte de las concordancias o las muestras elegidas por el profesor en un soporte distinto al programa de explotación de corpus (tanto en formato digital como en papel).

2.2.3 El impacto en el aprendizaje del enfoque DDL

Como dijimos anteriormente, este enfoque es considerado como efectivo para el aprendiente por su implicación directa en la construcción de su conocimiento de la lengua meta. Las razones para explicar esta efectividad son diversas, pero quizá la más importante es un cambio en la motivación del aprendiente, que se encuentra frente a una tarea desconocida, que le supone un reto y que rompe totalmente con los métodos más convencionales de enseñanza.

Más concretamente, y según el propio Johns (1991), el enfoque DDL comporta los siguientes efectos sobre el proceso de aprendizaje:

1. El uso del programa de concordancias tiene una influencia favorable estimulando la especulación y la necesidad de responder a ciertos interrogantes por parte del aprendiente, además de ayudarle al reconocimiento de patrones en la lengua meta y a extrapolar generalizaciones que expliquen su uso (como parte de un método inductivo).

2. Produce cambios en el papel del profesor, que pasar a ser director y coordinador de la investigación realizada de manera autónoma por los alumnos.

3. Se reconsidera el lugar que ocupa la gramática en el aprendizaje y en la enseñanza de lenguas. De un sistema que presenta al

alumno como un gestor de patrones y reglas que va aplicando en la construcción de textos en la lengua meta, se pasa a una visión mucho más centrada en las posibilidades del aprendiente y en su desarrollo psicolingüístico. En esta nueva visión, el alumno va forjando una conciencia gramatical gracias a los descubrimientos empíricos que realiza en muestras auténticas de uso de la lengua.

Aunque desde nuestro punto de vista el enfoque no carece de interés, no podemos dejar de considerar que existen ciertos *peligros* en su aplicación, sobre todo, en relación a las conclusiones o hipótesis extraídas de los datos por los alumnos. Puede ocurrir que se generalicen usos excepcionales de un determinado patrón lingüístico o elemento léxico, primando ciertas concordancias que no son tan generales, sólo porque aparecen con un mayor número de frecuencias en un corpus que consideramos representativo. No sugerimos con ello que se deba abandonar totalmente el uso de corpus, sino que deben usarse sin perder de vista que pueden proporcionar mucha información adicional sobre el uso de la lengua, complementando, contradiciendo o apoyando los recursos pedagógicos tradicionales y ayudando a crear y matizar lo que se denomina *conciencia lingüística*.

Cuando aprendemos, a la hora de interpretar los datos, nos sirven inevitablemente de guía los conocimientos anteriores y la intuición lingüística. Por otra parte, muchos de los posibles errores de interpretación pueden salvarse con una formación adecuada en el manejo de corpus y sus herramientas que permita entender cuáles son sus limitaciones y ventajas y profundizar en la metodología de búsqueda. En ocasiones, lo que ocurre no es que se utilice incorrectamente el programa de concordancias, sino que la metodología de búsqueda no es la adecuada, no se dan a los datos y resultados obtenidos la importancia que merecen o no se discrimina lo que es pertinente de lo que no lo es.

La exploración de corpus para la enseñanza, o *Data Driven Learning (DDL)*, aún no está muy extendida a los diferentes niveles educativos, debido a la dificultad de las tareas y soportes utilizados, así como a la escasez de dotación tecnológica y el insuficiente grado de alfabetización informacional. La bibliografía crítica y algunas de las investigaciones sobre esta metodología destacan que los resultados alcanzados son bastante limitados en alumnos con un nivel básico, por lo que suele

aplicarse, principalmente, en niveles avanzados y en entornos universitarios.

En general, muchos teóricos, entre ellos Aston (2001), consideran al DDL como un enfoque eficaz para desarrollar la conciencia metalingüística y metacognitiva de los aprendientes (Aston, 2001: 23), lo que les permite detectar las regularidades en los datos y por tanto interpretarlos. Por nuestra parte, consideramos que el DDL ofrece una visión de la lengua muy enriquecedora, puesto que expone al alumno a patrones, estructuras y vocabulario que no suelen encontrarse en los métodos comunes de aprendizaje de idiomas. Además, promueve una actividad de reflexión muy importante sobre la lengua meta y con ello, una mayor participación del aprendiente en su propio proceso de construcción del conocimiento.

Aunque habitualmente los corpus se han utilizado para estudiar distintas formas lingüísticas, podemos mencionar otros usos alternativos:

- La **práctica de la comunicación**, a través de tareas de resolución de problemas y de razonamiento y argumentación, en las que los aprendientes han de establecer un consenso sobre lo que están investigando.

- **Tareas relacionadas con el significado** de expresiones o palabras, utilizando lo que Bernardini (2001) ha dado en llamar el "principio de serendipia" (*serendipity principle*), es decir, la consulta del corpus de manera aleatoria, acercándose a él de forma desordenada, de forma que cada descubrimiento conduce a nuevos conocimientos.

- **Tareas relacionadas con la comprensión lectora**, ya que el enfoque inductivo necesita de una fluidez a la hora de leer e interpretar las concordancias. De hecho, Gavioli (Aston et al., 2004) señala en uno de sus estudios cómo el estudio de concordancias de palabras frecuentes de un texto puede ayudar a los aprendientes a predecir la estructura y el contenido del texto en su conjunto y a validar su conocimiento en léxico.

- **Tareas relacionadas con el léxico**. La exploración del corpus, incluso de manera no secuencial o a través de una búsqueda desordenada puede atraernos hacia expresiones o

palabras no conocidas, de las que investigaremos su significado y su uso, de forma que, poco a poco, se vayan incluyendo en el propio repositorio léxico del aprendiente.

- **Tareas relacionadas con el análisis textual y del discurso**. Normalmente se sigue un enfoque en el cual se analizan primero las estructuras, para después abordar el texto completo, pero podemos, realizar también la aproximación contraria, yendo del texto global hacia las concordancias. Esta aproximación ha sido bastante frecuente en estilística, donde las listas de frecuencias y las concordancias han servido para investigar los trabajos literarios de distintos autores, y observar así el tipo de léxico o campos semánticos propios de una determinada obra, aspectos culturales interesantes, ideologías, roles de personajes, etcétera. También se ha utilizado para comparar, en análisis contrastivos, textos de hablantes nativos con otros que no lo son, o entre nativos y aprendientes. Todo ello centrado tanto en la forma como en el significado, aunque poniendo el énfasis sobre este último. Finalmente, también se pueden realizar análisis intratextuales, en los cuales se buscan patrones en un texto que revelen ciertos significados de determinadas concordancias y del texto en su conjunto, o un análisis intertextual, en el que encontrar peculiaridades o regularidades de un texto en relación a otros.

- **Tareas relacionadas con la expresión escrita**. Los corpus son una herramienta de referencia para comprender el texto. Son, a su vez, otra forma de manuales de uso de determinados patrones y estructuras, que incluso superan a los tradicionales (diccionarios y gramáticas, entre otros) por el número de ejemplos que nos ofrecen y la variedad de usos que se pueden encontrar de un mismo ejemplo. De hecho, existen estudios como los de Bertacci y Aston (Cf. Aston, 2001) que muestran cómo los aprendientes pueden usar ciertos corpus a la hora de escribir o, incluso, realizar actividades como traducir para validar sus hipótesis, buscar el significado deseado o identificar las alternativas posibles para una estructura.

2.3 Compilación de corpus con fines pedagógicos.

La concepción y compilación de corpus con fines pedagógicos destaca la creación de corpus específicos, relacionados con el ámbito de la educación. La mayoría de los corpus a los que se alude en este campo son:

- los corpus implementados con textos procedentes de libros de referencia o manuales (*textbook corpora*),
- los corpus basados en la observación de la clase o en el aula,
- los corpus de aprendientes,
- los corpus de textos específicos de una determinada materia.

Dichos corpus sirven de apoyo para los dos enfoques anteriores, es decir, se implementan para que sean utilizados de forma directa o indirecta en el campo de la enseñanza.

Los **corpus con textos específicos** y vocabulario controlado son habituales en materias que necesitan un conocimiento muy específico de las estructuras y características de la lengua propia del campo de estudio al que aluden. Así hay asignaturas como francés con fines académicos, o inglés con fines específicos (por ejemplo, para biología, medicina, turismo, etcétera). También son muy apreciados en la enseñanza de la traducción, ya que son una herramienta muy útil para la consecución de una buena traducción o interpretación.

Los **corpus de aprendientes** y los **de observación en clase o en el aula** suelen ser utilizados en estudios sobre adquisición, de forma que se pueda conocer con exactitud cuáles son los procesos que sigue el aprendiente hasta alcanzar un determinado dominio de la lengua meta. Arrojan luz sobre las fases por las que pasa el aprendiente hasta la completa adquisición y las características de su lengua intermedia (o interlengua) en función del nivel de dominio en el que se encuentre. A través de estos estudios, hemos conocido qué aspectos lingüísticos se aprenden antes que otros, cuáles son más difíciles de aprender y las diferencias con los hablantes nativos, por medios de análisis contrastivos. Son, por lo tanto, una orientación muy útil para la creación de materiales y para el diseño curricular de asignaturas u otros cursos, pues nos

confrontan con las verdaderas características del aprendiente, sus necesidades y los obstáculos más habituales del proceso de aprendizaje[60].

Podemos concluir que el uso de corpus en la enseñanza es totalmente compatible con otras metodologías de enseñanza de lenguas como el enfoque comunicativo, al que podría servir como instrumento de apoyo para el desarrollo de destrezas que con otros recursos no son tan fáciles de abordar, pues no tiene la posibilidad de disponer de tal volumen de datos y de muestras de lengua diferentes de una forma rápida y sencilla.

Hay que decir que se desconocen por ahora investigaciones concluyentes que demuestren con claridad que el aprendizaje por descubrimiento como el que propone el *data-driven learning* –DDL-, sean más efectivo para desarrollar las competencias en lenguas que los métodos tradicionales basados en el estudio de la gramática y de las reglas. Se valora positivamente sus resultados, pero se desconoce su valor relativo en relación con el impacto de otros factores presentes en el proceso de aprendizaje como la motivación, el conocimiento de otras lenguas, la actitud o la dedicación.

No obstante, planteamos la conveniencia de promover estos enfoques entre los docentes. Para empezar, sería necesaria la difusión de los resultados obtenidos en foros de discusión y en encuentros entre expertos. Todo ello contribuiría a validar las conclusiones sobe su impacto real en el aprendizaje. Existe una abundante bibliografía al respecto, pero el número de publicaciones contrasta con la escasez de experiencias prácticas en las aulas. Por lo tanto, insistimos en la necesidad de colmar esa brecha para la que pueda aprovecharse el potencial del uso de los corpus en la enseñanza.

3. Principales ventajas de la aplicación de los corpus en la enseñanza de las lenguas

Los corpus, tanto orales como textuales, debido a la gran cantidad de datos que contienen sobre una lengua y debido a su formato digital, resultan de gran ayuda para la enseñanza de las lenguas (como lenguas

[60] Hablamos de los corpus de aprendientes en el primer capítulo del presente estudio y nos referiremos con más detalle a ellos posteriormente, en el capítulo 5 (p.117).

extranjeras, segundas o terceras lenguas, para fines específicos, con objetivos académicos...).

Como ya hemos indicado, el uso de corpus en enseñanza de segundas lenguas nos propone una nueva forma de enfocar el estudio de la lengua, en la medida en que consiste en un material auténtico procedente del uso real y no de un material artificial, subjetivo o guiado por la norma académica. Ante todo, los corpus han aportado una gran objetividad al conjunto de la investigación, ya que se basa en datos empíricos y no sólo en la introspección y la intuición de un determinado hablante. Ser un hablante nativo no implica automáticamente tener presentes todos los contextos de uso y las formas de una lengua. Es más, se sabe que no todos los nativos tienen las mismas intuiciones o percepciones sobre su propia lengua, lo que ha conducido, en algunas épocas, a hipótesis incorrectas y, por consiguiente, a reglas incompletas o inexactas sobre el funcionamiento de la lengua. El uso de corpus nos aporta, sin duda, una descripción superior de todos los fenómenos lingüísticos.

El uso de corpus, y en general de las TIC en la enseñanza, nos permite hoy en día incluir en poco espacio materiales muy numerosos y variados, así como una información que puede actualizarse rápidamente y que es fácilmente utilizable a través de una herramienta rápida, cuya consulta se puede dirigir en función de nuestros intereses. De hecho, las grandes innovaciones en el ámbito de las TIC han puesto a nuestra disposición muchos corpus en Internet gracias a lenguajes de marcado como XML o TEI (sin necesidad de recurrir al CD-ROM, que, no siempre estaba disponible, u obligaba a la adquisición de licencias de coste elevado y, por supuesto, sin perder ninguna de las funcionalidades de los anteriores corpus). Además, gozamos de una gran cantidad de herramientas de explotación *online* o en versiones instalables gratuitas. Entre otras posibilidades, podemos encontrar sistemas y programas de concordancias, etiquetadores y analizadores sintácticos, o programas de frecuencias o estadísticas necesarios para el análisis cuantitativo.

Los materiales de enseñanza de idiomas, aunque correctos y pedagógicamente adaptados al desarrollo de una lengua desde un enfoque convencional, contienen ciertas imperfecciones que podrían ser corregidas con el uso de corpus. En concreto, sabemos que muchos de los ejemplos que contienen son creados expresamente para cada aspecto gramatical, léxico o pragmático con que se quiera trabajar, por lo que, a

menudo, el alumno se encuentra con un vocabulario y unas estructuras poco frecuentes en contexto real, incluso artificiales. Además, existe una tendencia bastante elevada a la simplificación, es decir, se usa sólo una parte reducida del vocabulario existente, y ciertas estructuras rígidas que son reutilizadas en repetidas lecciones para garantizar su correcto aprendizaje. El cambio lingüístico o la variación no quedan reflejados. Esta simplificación, beneficiosa quizá en los primeros momentos de estudio de una nueva lengua, resulta, a la larga, muy reductora para el alumno, que no es confrontado a estructuras complejas con significados distintos, y a alteraciones habituales del orden más canónico y normativo de la gramática de una lengua (uno de los aspectos más característicos de la lengua hablada por los nativos y no la académica).

El análisis pormenorizado de un corpus oral o textual, nos ofrece otras visiones y usos de una lengua, que pueden ser incorporados a la enseñanza tradicional, enriqueciendo el aprendizaje y garantizando un espectro mayor de conocimiento al aprendiente. Incluso, se pueden introducir nuevas perspectivas no exploradas hasta ahora, como la variación lingüística, ya que, con el uso de corpus, el aprendiente puede beneficiarse del hecho de manejar datos recientes para poder observar los sucesivos cambios y el funcionamiento real de la lengua en un determinado contexto.

Los análisis de corpus pueden ofrecernos también aquellos usos más frecuentes de una lengua, y señalarnos incluso algunos de ellos que no son habitualmente recogidos por la gramática tradicional, y que pueden asimismo escapar a diccionarios u otras fuentes de consulta[61]. Esta modalidad resulta especialmente beneficioso para aquellos cursos que requieren un estudio más específico de la lengua, como es el caso de asignaturas de *lengua con fines específicos* (por ejemplo, francés para objetivos universitarios o FOU, francés aplicado al turismo, francés aplicado a la medicina, etcétera). Un estudio de corpus especializados podrá mostrarnos los aspectos más importantes del uso concreto de la lengua para este campo de conocimiento, facilitando el aprendizaje de

[61] Si bien un análisis de frecuencia de uso puede resultar interesante para el estudio de una lengua, algunos expertos como McEnery & Andrew (2001:120) recuerdan que la tradición ha criticado vehementemente este procedimiento, aduciendo que una limitación al estudio de los fenómenos lingüísticos más usuales puede contribuir a un empobrecimiento de los resultados de aprendizaje.

elementos realmente necesarios y comunes en él, garantizando una buena adecuación entre lo enseñado y su utilización en tareas futuras.

Evidentemente, el matiz de frecuencia de uso resulta controvertido en cuanto a selección de usos o de material que enseñar. Como sabemos, la concepción del o de los corpus que estemos manejando, la cobertura de la lengua que tenga, así como su representatividad y equilibrio, podrá privilegiar unos u otros aspectos morfosintácticos y léxicos. Es decir, dependiendo del espectro de contenidos que abarque el corpus, ofrece una determinada muestra de lengua, lo que puede crearnos una imagen de lengua que en realidad no es la más representativa o real. Si nos basamos sólo en este criterio para diseñar nuestro curso de lengua, es posible que privilegiemos aspectos que no son relevantes, sólo por que aparecen con más frecuencia en nuestro corpus. O al contrario, no abordar ciertos usos o estructuras porque no aparecen suficientemente en él. Sabemos que la frecuencia, sobre todo en corpus orales, es muy dependiente del contexto en el que aparece, por eso, a la hora de decidir qué hay que enseñar, los corpus han de ser una ayuda para tomar mejores decisiones, pero nunca han de utilizarse como respuesta definitiva.

Otra de las grandes ventajas que aportan los distintos enfoques de aplicación de corpus en la enseñanza, principalmente con el método directo (DDL), radica en que permiten al aprendiente adquirir muchos conocimientos por sí sólo, guiado o no, entroncando con lo que se conoce como autonomía en el aprendizaje, o también, como autonomía dirigida (si está siendo guiado por el profesor).

El uso directo de corpus, sobre todo de tipo oral, por parte de los aprendientes presenta numerosas ventajas, como la posibilidad de validar sus propias hipótesis sobre la lengua y el uso de esta, así como el descubrimiento y la observación de fenómenos particulares que no siempre se encuentran en los grandes corpus ni en los corpus escritos. Como señala Bernardini (Gavioli and Aston, 2001) un corpus de grandes dimensiones como el *British National Corpus* (BNC) es un terreno muy apropiado para fomentar el autoaprendizaje basado en la propia investigación, de forma que a partir de unas pautas iniciales, los aprendientes puedan investigar siguiendo sus propias ideas e intuiciones aspectos lingüísticos que sean de su interés, posiblemente por simple curiosidad enciclopédica. El profesor puede, en estos casos, actuar de asistente sugiriendo o ilustrando lo que los alumnos van investigando

relacionándolo con otros aspectos, actividades, y ayudando en la formulación de hipótesis finales a través del diálogo con los aprendientes. Se trata de llegar a una nueva metodología que, como señala Holec (1990), se base en la estrategia: observación, hipótesis, formación, uso. Boulton (2007:3) observa que los aprendientes pueden ver tendencias en el uso o en las concordancias mostradas, incluso si no son capaces de verbalizarlas. En un principio, sus observaciones serán muy generales, pero en función de su progreso en el conocimiento de la lengua, se irán afinando.

El uso directo de corpus requiere un esfuerzo complejo, pero ofrece numerosas ventajas. El introducir este enfoque de forma progresiva podrá, sin duda, mejorar la calidad de la enseñanza, ya que se fomentará el desarrollo de competencias y estrategias en el aprendiente que no podrían ser abordadas de otra forma. Insistimos en la necesidad de su utilización en el aprendizaje de lenguas, como complemento, en un principio, a los métodos más convencionales, ya que contribuirán a dotar al aprendiente de una lengua meta mucho adecuada a los contextos de uso, algo muy beneficioso para los resultados de aprendizaje.

3.1 Data-Driven Learning –DDL- y sus beneficios en el aprendiente

El tipo de metodología empleada en el desarrollo de actividades o cursos con corpus, como ya queda explicado, se centra en el aprendiente, el cual ha de reflexionar, indagar y analizar los datos disponibles para llegar a sus propias hipótesis.

De hecho, Boulton (2010a), después de investigar la metodología de estudio y uso de corpus con alumnos de distintos estilos de aprendizaje[62], concluye inicialmente que podría convenir a muchos de los distintos aprendientes, puesto que cada uno abordaría las tareas a través de su estilo predominante. De lo que se deduce que una característica destacable de este enfoque es que la consulta directa de corpus en el

[62] Para definir los estilos básicos se nutre del "Index of Learning Styles" de Felder y Silverman: R.M. Felder and L. K. Silverman (1988): "Learning and Teaching Styles in Engineering Education", en: *Engineering Education*, 78 (7), pp. 674-681.

ordenador permite que cada estudiante pueda investigar o trabajar según sus propias necesidades, sus intereses o dudas, sin necesidad de seguir el mismo patrón de enseñanza o recorrido generalista de una metodología de enseñanza que no tiene en cuenta las individualidades. El enfoque implica al alumno de manera activa, pues es más consciente de su desarrollo cognitivo y de los procesos en desarrollo (aprendizaje significativo), lo que le lleva a fijar más fácilmente sus conocimientos.

Además de explotar el espíritu crítico del aprendiente (sobre todo en el caso del método inductivo), este tipo de actividades fomenta su participación, la argumentación, la defensa de sus propias ideas, por lo que es una excelente manera de garantizar la interacción (generalmente oral) de los aprendientes en la lengua de estudio. A menudo, cuando el trabajo no es dirigido, sino que se basa en el autoaprendizaje, da lugar a distintas interpretaciones, análisis de distintas concordancias, lo que conlleva que los aprendientes tengan que negociar sus hallazgos, discutir sus análisis para llegar a una puesta en común de sus conclusiones. No sólo se despierta la conciencia lingüística del aprendiente de L2, sino que son más conscientes de aspectos gramaticales, léxicos o textuales. Se les permite crear así sus propios criterios descriptivos y un marco formal para sus hipótesis sobre la lengua, siendo capaces, poco a poco, de cuestionar aquellas de sus profesores, de los libros de texto y de otros manuales de referencia que puedan utilizar.

Supone, por tanto, una excelente alternativa a los métodos pedagógicos convencionales, que, generalmente, están basados en una muestra de lengua artificial. De hecho, existen estudios que demuestran que los fenómenos lingüísticos aprendidos de forma aislada, a través de lecciones de gramática clásica, sólo son movilizados por los alumnos en contextos similares (por ejemplo, en un examen o en la propia clase de lengua), pero que difícilmente serán utilizados en otros contextos con una intención o finalidad comunicativa distinta. El trabajo con corpus, por consiguiente, mejora el uso de la lengua correspondiente a un contexto de comunicación estereotipado y posibilita una mejor capacidad de adaptación a contextos variados que exceden el marco del entorno académico.

En muchas ocasiones, los aprendientes realizan sus adquisiciones guiados por el profesor. La mayoría de las veces este aprendizaje en contexto educativo se produce en un país donde la lengua de estudio no se habla (es decir, en un contexto puramente exógeno). Por tanto, las

posibilidades de que el aprendiente pueda, ya no sólo comunicar, sino escuchar la lengua meta fuera del aula son bastante limitadas.

Si bien la comunicación con nativos de la lengua no es algo que nos pueda facilitar un corpus, en cambio, gracias a los corpus podremos "recrear" la situación de comunicación más cercana a la de un medio endógeno (donde se encontrarían en inmersión lingüística), sobre todo si se trata de un corpus oral de nativos, en el que podamos encontrar conversaciones en entorno no formal (familiar o de ámbito privado). Con ello, tal vez no mejoremos la expresión oral, pero es posible que ayudemos al aprendiente a conocer muchos de los fenómenos presentes en la lengua oral, a acercarse a la llamada "gramática de la lengua oral", a aspectos pragmáticos habituales (como la cortesía, negociación de los turnos de habla, la atenuación, organización del discurso…) y, por consiguiente, a mejorar su capacidad de comprensión oral.

El aprendiente asume así la complejidad de la lengua que aprende y es capaz de analizar y comprender estructuras y construcciones complejas, pero que son, sin embargo, fáciles de encontrar en situaciones comunicativas corrientes.

Por otro lado, sabemos que las ventajas serían mayores si el aprendiente no se dedicara sólo a reunir datos estudiados en el corpus o a verificar hipótesis en ellos. Sería mucho más provechoso para el aprendiente si pudiera constituir y crear su propio corpus (entendiendo crear no sólo como realizar las grabaciones pertinentes, sino también diseñar, transcribir y anotar corpus).

Para un aprendiente, tener la posibilidad de crear o implementar su propio corpus, supone un ejercicio de aprendizaje de una lengua inigualable. El hecho de crear su propio corpus, la implicación a la hora de recogida de datos y sobre todo, la transcripción, resulta un ejercicio muy beneficioso tanto para la mejora de la comprensión oral como para la expresión escrita.

La transcripción entronca así con una serie de ejercicios clásicos muy demandados en los distintos métodos de instrucción como dictados (discriminación fonética), ejercicios en los que hay que rellenar huecos basándonos en lo comprendido, o bien, en las actividades de reescritura. En este caso, el alumno emprende una nueva forma de abordar ambas destrezas, que aumentará su motivación, y que le atraerá nuevas dudas o problemas distintos a los habituales, que tendrá que solucionar,

provocando un grado de asimilación mayor gracias a la implicación en la tarea.

Uno de los aspectos más llamativos del enfoque es el análisis particular de los datos, puesto que el alumno deberá ser consciente de todo lo dicho para poder transcribirlo tal y como aparece en la grabación (incluidos algunos errores o incoherencias). Ello le llevará también a distinguir entre lo correcto y lo incorrecto, dándose cuenta de los errores y haciendo un esfuerzo por averiguar su naturaleza e iniciando distintas reflexiones sobre la lengua y sobre su propia producción oral.

4. Críticas frecuentes a la aplicación de corpus en la enseñanza

Pese a que la aplicación de corpus en enseñanza de lenguas ha tenido bastante aceptación, sobre todo en investigadores de lingüística aplicada y en entornos universitarios, no se libra de algunas críticas esenciales y de ciertos detractores.

Principalmente, han sido objeto de muchas críticas por parte de algunos investigadores, que los consideraban como poco representativos de una lengua, incapaces de recoger en lo fundamental las diversas formas de uso e insuficientes para presentar conclusiones generales sobre las características y rasgos de una lengua.

Otros autores son de la opinión de que un corpus contiene una cantidad enorme de conocimiento previo, tanto a nivel lingüístico como cultural, que puede resultar perjudicial para algunos aprendientes, ya que puede confundirles en sus análisis y en sus interpretaciones (Aston 1997).

Como venimos señalando, es muy importante una buena elección del corpus para que este tipo de actividades resulten beneficiosas, y una adecuada adaptación al nivel de nuestros aprendientes. En este punto, podemos afirmar también que quizá un corpus de aprendientes pueda ser beneficioso para usuarios no iniciados en la metodología y que, además, posean un nivel aún básico de lengua, puesto que reflejará una realidad que seguramente les resulte familiar, y en la que además, pueden participar, aportando otras respuestas, y detectando posibles errores (que están a su alcance).

Muchos autores consideran que los corpus pueden proporcionarnos una visión y una descripción muy clara de la realidad de una lengua. Mientras, otros se preguntan si ese es el tipo de lengua que necesitan conocer los aprendientes no nativos (Gavioli and Aston, 2001: 238). Es cierto que la extensión de los corpus que conocemos (incluso para el inglés, que ha sido la lengua que goza de corpus de mayor tamaño) es mucho menor que la media de estructuras que puede reflejar la competencia y la experiencia de cualquier adulto en el uso cuotidiano de su lengua. Un corpus no podrá abarcar la totalidad de los usos de una lengua, sin embargo, no podemos dudar de su utilidad para señalar los usos más frecuentes y, por tanto, para atribuir a las muestras de lengua criterios de pertinencia para la enseñanza.

En muchas ocasiones, también el hecho de no encontrar ciertas estructuras, y por tanto, no poder validar empíricamente alguna de las hipótesis tradicionalmente aprendidas, puede ser positivo. La ausencia de resultados de búsqueda puede llevar a un análisis más profundo, a nuevas búsquedas hasta hallar el porqué, y por consiguiente, al realizar un trabajo más reflexivo, es posible que se dé un aprendizaje más profundo y que dure por más tiempo indexado en la memoria a largo plazo.

Sabemos que los corpus de referencia son colecciones de textos orales o escritos almacenados digitalmente que han sido producidos, a menudo, por hablantes nativos. Muchos autores, entre ellos Cook y Carter (Gavioli and Aston, *op.cit*), señalan si es necesario que los aprendientes necesiten imitar el comportamiento y la lengua de los hablantes nativos, y si por tanto, los corpus son modelos relevantes de ello. Esto no incluye sólo a los documentos procedentes de corpus, sino también a los documentos auténticos en general.

En este punto volvemos a la idea de la representatividad de un corpus y al grado de aproximación a la realidad de una lengua. Pese a que es habitual encontrar cada vez corpus más desarrollados y completos, que reflejan gran número de situaciones comunicativas distintas, a la hora de mostrar determinados modelos de conversación habitual (como parte, por ejemplo, de un enfoque comunicativo), es posible que no encontremos lo que necesitamos. En estos casos, a veces es mejor recurrir a documentos auténticos o bien, a inventar un diálogo verosímil que pueda actuar como modelo. Un corpus puede servir entonces para comparar distintos modelos, puesto que la realidad de los datos contenidos en un corpus tiene que servir, principalmente, para conocer e

interpretar la lengua y que el aprendiente/hablante cree sus propios modelos a partir de él. Como señalan Gavioli y Aston (*op.cit*), no se trata de que los aprendientes imiten aquello que vean, encuentran o estudian en los corpus, sino que asimilen el conocimiento que ofrecen y produzcan, a partir de él, su propio discurso.

Como hemos citado anteriormente, una de las principales desventajas que los docentes ven en los corpus es su complejidad, no sólo por su forma y por su contenido, sino también por las herramientas informáticas que se han de conocer para su explotación. De hecho, muchos de los corpus no se han implementado teniendo en mente una posible aplicación pedagógica, sino más bien los requerimientos de una determinada investigación, lo que contribuye a que se alejen enormemente de los posibles usos con aprendientes.

Por consiguiente, su utilización no se ha extendido lo suficiente, quedando reducida al ámbito universitario, lo que ha alimentado la creencia de que sólo son apropiados para un nivel avanzado de lengua y en contexto educativo superior (Boulton, 2008). Volvemos en este punto a comentar que esta idea no ha sido suficientemente demostrada, puesto que no existe un número importante de análisis empíricos con aprendientes de niveles no universitarios que puedan avalar esta teoría[63].

Sabine Braun (2007) señala que algunos de los problemas para la inserción de corpus en la educación secundaria provienen no sólo de su complejidad o de la inexperiencia de los aprendientes, sino también de las necesidades curriculares o el nivel de consecución de objetivos concretos que ha de alcanzarse en los distintos cursos. Un currículo que ha de adaptarse, además, al conjunto del programa educativo para el curso, es decir, a unos requerimientos globales. Muchos docentes piensan que los corpus sólo pueden ser actividades complementarias, ya que no existe un lugar para ellos dentro de un desarrollo curricular aparentemente tan rígido.

Por tanto, no podemos obviar que los corpus son complejos, y que constituyen un salto de nivel considerable para los aprendientes, acostumbrados a libros y manuales con una lengua más ordenada, sencilla y transparente. Sin embargo, existen muchas iniciativas posibles

[63] Muchos de los estudios se realizan con aprendientes universitarios porque los investigadores dependen de su entorno más inmediato para llevarlos a cabo más fácilmente.

para salvar estos obstáculos. Para evitar la complejidad, tenemos que acudir a una revisión de la metodología del profesor, que tendrá que ayudar a los alumnos en su utilización de corpus, realizando desde el principio una introducción a estos y a las herramientas de manejo y a una pre-edición de los corpus con los que trabajar. Existen muchos caminos para lograrlo, entre los que destacan el utilizar primero corpus más sencillos, o una selección de muestras donde sean más fácilmente identificables los patrones que queremos subrayar. También se puede trabajar con corpus de forma gradual, es decir, comenzado con colecciones de textos pequeñas con forma similar, para pasar después a las más ricas y variadas, o incluso, realizar cambios en las transcripciones, suprimiendo elementos que, en un principio, puedan desorientar a nuestros aprendientes (Como bien señalan autores como Boulton, 2009a; Römer, 2006). Es evidente que el uso de corpus supone un cambio en la metodología de trabajo, algo que, *a priori*, desalienta a muchos posibles usuarios.

Para los docentes, a diferencia de diccionarios, manuales de referencia y otros textos de consulta, los corpus no pueden solventar las dudas o necesidades de forma directa e inmediata que los aprendientes tienen en el transcurso del aprendizaje y en momentos concretos de uso de la lengua. Esta crítica reside en la forma tradicional que aún tenemos de ver la apropiación de una lengua, donde se cree que los conceptos se construyen con la memorización y que no valora el análisis de la lengua y el desafío de la construcción de una conciencia lingüística como algo útil, por ser laborioso y poco efectivo a corto plazo. Muchos aducen que no es necesario volver a inventar las reglas si ya están escritas, obviando teorías psicolingüísticas que sustentan que las reglas, precisamente por ser construcciones artificiales externas al contexto de aprendizaje, suponen un mayor esfuerzo cognitivo para el aprendiente.

Aunque la mayoría de los corpus utilizados en la enseñanza de lenguas con los que se han realizado estudios son grandes corpus textuales de referencia, con un tamaño suficiente para ser tildados de representativos, se observa que el recelo hacia el uso de corpus orales en ese campo es aún mayor. Primero, por su escasa disponibilidad que tiene que ver con la dificultad de su implementación. Y segundo, por su forma, mucho más compleja, al incluir en su transcripción numerosos símbolos concretos para reinicios, reformulaciones, pausas prosódicas, etcétera, con los que el aprendiente no está familiarizado y que no facilitan su lectura.

Por otra parte, los corpus orales tienen que lidiar con los prejuicios frente a la lengua oral que todavía muchos consideran demasiado familiar, y por tanto indigna de ser convertida en objeto de enseñanza.

Como venimos señalando, el problema principal radica en el gran desconocimiento de los corpus por parte del cuerpo docente. Se desconoce su potencial de aprendizaje, y se infravalora a la lengua oral como base para el fomento de la capacidad de comunicación del estudiante. El uso de corpus no puede ser considerado como la panacea, como afirma Widdowson (2000), uno de los teóricos de la Lingüística más críticos con los corpus[64], pero resultan un apoyo muy útil si sus resultados se utilizan con cierto espíritu crítico. A nuestro entender, deben mantenerse siempre ciertas reservas frente a la tendencia a generalizar el éxito de ciertos resultados, pero sin dejar de valorar la contribución de los corpus a la motivación y la interiorización de estructuras y usos variados de la lengua.

5. Panorama actual de la aplicación de los corpus en enseñanza de segundas lenguas

Existen iniciativas interesantes por parte de investigadores generalmente individuales que promueven el uso de corpus fundamentalmente en el entorno universitario. Sin embargo, estimamos que los resultados de estos estudios no son suficientemente esclarecedores de la incidencia real del uso de corpus en el aprendizaje. Ello incide, como ya queda explicado en la escasa difusión del enfoque en la enseñanza.

Queremos incidir aquí en el gran desfase entre la investigación realizada en la Lingüística de corpus y la transferencia de sus resultados en la enseñanza de las lenguas. Existe una gran diversidad de propuestas de aplicación, sugerencias de uso y estudios empíricos con alumnos de todo nivel, reflejadas en artículos y libros esenciales de la bibliografía de este campo. Sin embargo, la mayor parte de ese conocimiento no va más allá del planteamiento teórico.

---

64 Algunas de sus críticas se han señalado en el primer capítulo del presente estudio, entre ellas, destaca la de los problemas de contextualización de los textos y la representatividad.

La lengua en la que más se han utilizado los corpus aplicados a la enseñanza es, como era de esperar, el inglés, en torno al cual abundan no sólo los estudios científicos[65], revistas e incluso, conferencias especializadas, sino una copiosa producción de material de referencia basado en corpus como diccionarios o libros de texto, herramientas informáticas o aplicaciones CALL.

Este uso avanzado de corpus en enseñanza de lenguas se ha extendido tanto en los Estados Unidos como en algunos países de Europa como Bélgica (destacando el trabajo de la profesora Sylviane Granger), Reino Unido, Italia (con la investigaciones de Guy Aston y Silvia Bernardini, muchas para el campo de la traducción) o Francia.

En España, el uso de corpus está poco extendido, pese a algunas experiencias interesantes, y nada parece indicar que pueda haber una expansión significativa en los próximos años, pues es previsible que la introducción de procedimientos innovadores de este tipo corra una suerte paralela a la expansión de las TIC en la enseñanza, que como todo el mundo sabe, se está enfrentando a multitud de resistencias.

A lo largo del presente apartado, describiremos algunas de las aplicaciones más representativas que han tenido lugar en Francia y en España, por ser el ámbito de estudio más cercano al de nuestra investigación.

5.1 Uso de corpus en adquisición de segundas lenguas en Francia

Francia es uno de los países europeos más consciente de las ventajas que supone el uso de los corpus para el aprendizaje del francés, pese a haberlo hecho algo tarde con respecto al inglés. Dichas iniciativas son impulsadas por el Ministerio de Educación, directamente comprometido con la sistematización y la creación de corpus.

De modo que existen numerosos grupos de investigación especializados que están desarrollando aplicaciones muy interesantes, aunque aún poco accesibles para el gran público. No en vano, Francia

[65] Boulton (2008:40) realiza un estudio sobre la aplicación del método DDL en Europa, concluyendo que la lengua más estudiada era el inglés (33 de 39 estudios) y su ámbito de aplicación más frecuente, los hablantes de L2 adultos, mayoritariamente en un contexto académico de enseñanza superior.

cuenta con una exhaustiva clasificación de los distintos corpus realizados por sus investigadores, promovida por la *Délegation Génerale à la Langue Française et aux Langues de France* (DGLFL) y una guía de buenas prácticas para la realización de corpus: *Guide des bonnes pratiques*, publicada por el CNRS (*Centre National de Recherche Scientifique*).

También cuenta el país vecino con una gran base textual de consulta, FRANTEXT, con más de 200 millones de palabras, sin embargo, es sorprendente que no cuente con un gran corpus oral de referencia. Se han producido varios intentos de crear corpus orales más o menos representativos, pero no son lo suficientemente conocidos ni explotados. Quizá por ello la utilización en Francia de corpus en enseñanza se restringe más al género textual, y por supuesto, como en otros países de Europa, a otras lenguas como el inglés como lengua extranjera.

Pese a esta limitación, es de destacar que los franceses fueron los pioneros, casi sin pretenderlo, en el uso de corpus en la enseñanza, con el método *Les Orléanais ont la parole* (1976), creado en el ámbito de desarrollo e implementación de uno de los primeros corpus orales de Francia, el corpus ESLO[66], en su primera fase.

A partir de ahí han surgido distintos grupos de investigación que se han ocupado de analizar el impacto del uso de corpus en la enseñanza de lenguas, con diferentes proyectos y herramientas. En este ámbito, se podrían citar decenas de ellos, pero destacaremos como más representativos por su producción y larga trayectoria, a los grupos TALEP, antiguo GARS y DELIC, de la Universidad de Provence, CRAPEL de la Universidad de Nancy, y LIDILEM de la Universidad de Grenoble.

Llegados a este punto, una vez más debemos lamentar que ocurra lo mismo que para otras lenguas y otros países: el gran desfase entre la producción científica en torno al tema y la aplicación real de sus resultados a las prácticas de enseñanza.

Son frecuentes los estudios que intentan demostrar la utilidad del uso directo de corpus o *data-driven learning* con estudiantes universitarios, como los llevados a cabo por Boulton (2007, 2008, 2009a, 2009b, 2010a, 2010b, 2011) en el seno del centro CRAPEL. En este caso, Alex Boulton

[66] Corpus descrito en el primer capítulo del presente estudio.

realiza sus investigaciones en el ámbito del inglés con fines específicos. El estudio consiste en examinar la capacidad de los hablantes francófonos aprendientes de inglés, para trabajar algunos aspectos del DDL sin un entrenamiento previo, y en distintos niveles (incluido el más básico). También se ocupa de investigar la adquisición de algunos patrones o estructuras en estudiantes de ingeniería que cursan la asignatura de Inglés, siguiendo la metodología DDL (como los tiempos verbales *will* y *going to* -Boulton, 2007). En otros de sus estudios (Boulton, 2010a) analiza los estilos de aprendizaje que resultan reforzados por el uso de la metodología DDL. Finalmente, Boulton realiza estudios comparativos entre la enseñanza convencional y el enfoque basado en corpus en relación con ciertas áreas problemáticas de la adquisición[67], que habían sido previamente identificadas con tests de lengua (Boulton, 2010b).

En ninguna de sus investigaciones, y así no lo expresa el propio Boulton, se ha logrado demostrar el éxito relativo del uso de corpus para la enseñanza de la lengua. El resultado más positivo al que se ha llegado es que con esta metodología se pueden alcanzar el mismo nivel de resultados de aprendizaje que con un método convencional. Por el contrario, se obtienen resultados superiores en cuanto a conciencia de la lengua, puesto que se sabe que tras el uso de los corpus los aprendientes son capaces de detectar los patrones en las concordancias de muestras reales de lengua y aplicarlas adecuadamente posteriormente en nuevos contextos como lo harían con las reglas gramaticales a las que están acostumbrados (Boulton, 2007: 10).

Como ejemplo, el centro CRAPEL está llevando a cabo en la actualidad el proyecto FLEURON[68] (*Français Langue Étrangère*

[67] Concretamente, se analizan 15 ítems tradicionalmente considerados como áreas problemáticas de adquisición para los aprendientes francófonos. Suelen enseñarse desde niveles básicos, pero raramente llegan a asimilarse correctamente. Estos ítems se han seleccionado previamente de las producciones escritas de los aprendientes implicados en el estudio, focalizándose en aspectos de gramática y uso, y no en problemas de ortografía o semántica. Para el desarrollo del estudio, con un grupo se trabaja con metodología DDL sobre el *British National Corpus*, pero con apoyo de material impreso y con actividades ya concebidas, lo que permite focalizarse sobre el elemento que se quiere trabajar y evita las distracciones derivadas por la abundancia de datos que resulta del sistema de búsqueda de concordancias. Otro grupo trabaja normalmente con metodología tradicional.

[68] Apenas encontramos difusión de este proyecto, pero podemos ver sus directrices aquí: <http://www.msh-lorraine.fr/index.php?id=353 >

Universitaire: Ressources et Outils Numériques), que comenzó en el año 2005, se interrumpió, pero ha sido retomado a principios de 2009. Su finalidad es la creación de una base de recursos destinado a los estudiantes extranjeros que quieran mejorar su nivel de francés antes o durante la estancia en una universidad francesa.

Esta interfaz de ayuda al estudiante presenta una serie de situaciones comunicativas frecuentes a las que son susceptibles de estar confrontados los estudiantes una vez en Francia. El *sitio* pretende aunar un espacio tanto para el desarrollo de las habilidades lingüísticas, como de las estrategias o capacidades de aprendizaje, que se pueden utilizar simultáneamente. Se ofrecen así recursos lingüísticos y pedagógicos, a saber, una gran cantidad de documentos auténticos (unas 400.000 palabras divididas en entrevistas, conversaciones y narraciones con unas 75 horas de duración) que muestran las situaciones comunicativas elegidas junto con actividades de aprendizaje relativas a la comprensión y a la expresión, así como fichas descriptivas precisas de éstas catalogadas informáticamente. Se trata de conciliar los aportes del enfoque comunicativo con los del enfoque *form-focused,* o centrado en la forma, habilitando la observación a comportamientos lingüísticos reales y efectivos. Continúa en desarrollo y se prevé alcance las 700.000 palabras y una centena de horas de grabaciones.

Otro proyecto que se apoya en el uso de corpus, en este caso de aprendientes, lo encontramos en *Freetext*[69], que pretende desarrollar un programa de enseñanza de lenguas asistida por ordenador (ELAO) para trabajo con aprendientes desde niveles intermedios hasta avanzados. Se compone de cuatro tutoriales, con 16 documentos auténticos, con texto y fichas audiovisuales, que describen distintos actos comunicativos y una serie de actividades de explotación. Dichas actividades se basan en el corpus FRIDA[70], corpus de aprendientes del francés, utilizado en este proyecto para hacer un recuento de los errores más frecuentes del público potencial de la herramienta y poder concebir así las actividades más adecuadas.

Además de estos proyectos donde los corpus suponen un aporte esencial, podemos encontrar la colaboración de diversos grupos para fomentar el uso de los corpus de aprendientes por medio de distintas

[69] Encontramos información del proyecto en: http://www.latl.unige.ch/freetext/en/description.html
[70] http://www.uclouvain.be/en-cecl-frida.html

herramientas de explotación. Sabemos que la lengua del aprendiente supone un reto para la mayoría de los programas informáticos existentes, que no son capaces de desarrollar sus tareas de forma efectiva debido a las características especiales de esta.

En este ámbito, en Francia, encontramos el caso de la red *Kaleidoscope*[71], que agrupa a 24 países y 91 grupos de investigación, entre ellos, uno de los ejes del laboratorio LIDILEM de la Universidad Stendhal (Grenoble, Francia). En concreto, el eje de investigación del LIDILEM especializado en tratamiento automático de la lengua, representado por G. Antoniadis, O. Kraif, C. Ponton y V. Zampa, proporciona la herramienta *Exxelant* (*Example eXtractor Engine for LANguage Teaching*)[72]. Esta aplicación web, realizada a partir del corpus FRIDA (*French Interlanguage Database*, del que hablaremos posteriormente en este estudio), pretende ser un sistema capaz de realizar búsquedas y análisis de los errores anotados y categorizados en el corpus. Aparentemente es una herramienta que ha logrado bastante efectividad en su aplicación a FRIDA, pero que no ha sido difundida, quedando enmarcada dentro del proyecto en el que nació y sin posibilidad de ser consultada de forma libre. De hecho, creemos que no ha sido probada con otros corpus de aprendientes, ya que necesita una transcripción muy concreta, con una serie de símbolos y etiquetas determinadas para los errores, lo que complica su reutilización.

El grupo LIDILEM también es el encargado del sistema MIRTO (*Multi-apprentissages Interactifs par des Recherches sur des Textes et l' Oral*), aún en desarrollo, y que pretende ser una plataforma de enseñanza de lenguas de uso sencillo, que, ayudándose de las técnicas propias del tratamiento automático de lenguas, ponga a disposición de los docentes una serie de recursos y herramientas informáticas que les ayuden en la concepción de actividades y escenarios pedagógicos basados en corpus escritos. Además, pretenden la inserción de estos escenarios en sistemas de aprendizaje a distancia, de manera que el aprendiente obtenga del sistema un cálculo de su nivel, la posibilidad de explicar sus errores durante el test de nivel y de ofrecerle una adaptación de los escenarios pedagógicos

[71] Información de la red en: http://www.noe-kaleidoscope.org/telearc/

[72] Debemos señalar que algunos investigadores como Olivier Kraif han realizado también otra serie de aplicaciones informáticas para la explotación de corpus, como el sistema de búsqueda de concordancias para expresiones complejas *ConcQuest*, o el programa *Alinéa*, para la implementación y edición de corpus bilingües alineados. (http://w3.u-grenoble3.fr/kraif/).

en función de sus conocimientos y su nivel de competencia. (Antoniadis et Ponton, 2004: 4). En la actualidad, parece que sólo está desarrollada la parte dedicada a la concepción de ejercicios, aunque no existen muchos datos al respecto, ni ninguna versión disponible para su prueba.

Como podemos observar, si bien existen numerosas tentativas de utilizar los corpus en la enseñanza, sobre todo a partir de herramientas de explotación, muchas se encuentran aún en desarrollo o no han conseguido las expectativas deseadas, por lo que no han sido difundidas. Evidentemente, dada la complejidad de este tipo de aplicaciones, es necesaria una mayor colaboración entre expertos en lingüística de corpus, en lingüística computacional y en tratamiento automático del lenguaje para llevar a cabo nuevas herramientas y programas. Sin duda, la posibilidad de contar con sistemas de explotación sencillos (sin contar otros programas como *Wordsmiths Tools*), que permitieran además desarrollar actividades, escenarios pedagógicos u otras tareas propias de la didáctica de lenguas, animaría a los docentes a utilizar los corpus, y por tanto, a que estos se desarrollaran más.

5.2 Corpus y su uso en adquisición de segundas lenguas en España

El panorama del uso de corpus para la enseñanza de segundas lenguas en España es aún menos alentador. En la actualidad, son pocos los docentes que se ocupan de su estudio y aplicación, y como hemos dicho anteriormente, apenas aparecen fuera del ámbito universitario.

Un gran paso para su difusión se ha dado con la creación de la asociación AELINCO[73] (Asociación Española de Lingüística de Corpus), que, en sus congresos anuales, recoge muchas de las iniciativas que se están realizando a nivel nacional en cuanto a la implementación y uso de corpus se refiere.

Actualmente en España las investigaciones sobre corpus en la enseñanza se realizan principalmente con dos idiomas: Inglés como lengua extranjera o con fines específicos y Español como lengua

[73] http://www.um.es/aelinco/

extranjera. Pocos o prácticamente inexistentes son los estudios sobre el Francés.

En Inglés como lengua extranjera mencionaremos los trabajos de Cantos, Sánchez y Criado en la Universidad de Murcia, ciertas investigaciones para la enseñanza de patrones sintácticos con alumnos del Grado de Lengua Inglesa en la Universidad Politécnica de Valencia a través de una instrucción basada en corpus (*corpus-based*) para aprendizaje de Lexicografía, las colaboraciones de Amaya Mendicoetxea (UAM) para el corpus ICLE (UAM-ICLE) y las investigaciones sobre análisis de errores semiautomático de un corpus de aprendientes de inglés L2 textual por Díez Belmar en la Universidad de Jaén.

En el ámbito del Español como lengua extranjera destacamos las investigaciones avanzadas del Laboratorio de Lingüística Informática de la UAM, centro pionero sobre la implementación y tratamiento de corpus orales de toda índole. En la actualidad, trabajan en la implementación de un corpus de aprendientes de ELE provenientes de distintos ámbitos geográficos y con lenguas maternas diversas y en su explotación a través del análisis de errores. También han utilizado el corpus oral de referencia C-ORAL-ROM para su aplicación en el aula, para lo que han desarrollado un estudio de los textos orales del corpus que han implementado después en materiales de apoyo para el trabajo de la destreza de comprensión auditiva (que será expuesto posteriormente en el presente capítulo).

Otros ejemplos lo constituyen los trabajos de Amaya Mendicoetxea (UAM) y Cristóbal Lozano (Universidad de Granada) para la realización de un corpus escrito del español como L2 de hablantes con lengua materna inglesa (CEDEL-2[74]).

Para el francés como lengua extranjera, fuera del presente estudio, no conocemos trabajos con corpus que resulten significativos o que continúen en desarrollo en la actualidad. Existe una amplia tradición investigadora sobre el uso de las TIC en la enseñanza del francés[75], pero

[74] http://www.uam.es/proyectosinv/woslac/cedel2.htm

[75] Destacan los trabajos para el uso de la Web 2.0. de Carmen Vera, los itinerarios pedagógicos y el campus virtual de la Universidad de León, coordinados por Mario Tomé y los proyectos con corpus como el Giapel de la Universidad de Castellón, coordinados por Mª Luisa Villanueva. También podemos citar aquí proyectos de comunicación a través del ordenador (*Computer Mediated Communication*) llevados a cabo por la Universidad Complutense a través del proyecto Galatea, y algunos proyectos de

los corpus han sido un área poco explorada por el momento. En su mayoría, las aplicaciones se centran en corpus textuales específicos, muchos de ellos de obras literarias, que son utilizados de forma puntual para la enseñanza de la literatura francesa.

Por tanto, podemos afirmar que nuestro estudio retoma el interés surgido para la enseñanza de otras lenguas, ya que no existe en su ámbito ningún otro de sus mismas características, sobre todo, por contar con aprendientes y por su carácter oral y espontáneo.

5.3 El proyecto C-ORAL-ROM-ELE del LLI-UAM: C-ORAL-ROM en la enseñanza de Español Lengua Extranjera (ELE)[76]

El antecedente principal de esta tesis es el trabajo realizado en el Laboratorio de Lingüística Informática de la Universidad Autónoma de Madrid. Este trabajo pretende utilizar el corpus de habla espontánea C-ORAL-ROM para la didáctica de español para extranjeros, permitiendo sobre todo trabajar la compresión auditiva junto con ciertos contenidos gramaticales asociados para algunos niveles concretos de lengua enmarcados dentro del Marco común europeo de Referencia para las lenguas (MCER).

En un primer momento se realizó un análisis y clasificación de los textos orales del corpus por niveles, encuadrándolos en distintas funciones y nociones comunicativas y aspectos gramaticales interesantes. Posteriormente, este trabajo previo se ha implementado en un conjunto de actividades para la práctica de la comprensión auditiva y del uso de la lengua en contexto.

La fase inicial del proyecto "integra el uso de corpus lingüísticos para la didáctica (...); y al mismo tiempo la aplicación de programas de enseñanza de lenguas asistido por ordenador (ELAO), pues está construido en una plataforma informática que permite realizar análisis

enseñanza de la didáctica y la intercomprensión de lenguas romances en una plataforma europea como Galapro de la Universidad Autónoma de Barcelona.

[76] Este trabajo previo se ha materializado en el libro: CAMPILLOS LLANOS, L., GOZALO GÓMEZ, P., GUIRAO MIRAS, J. Mª, MORENO SANDOVAL, A. (2010): *Español oral en contexto. Vol. 1. Textos de español oral. Material de ELE basado en corpus. Comprensión auditiva*. Madrid: Servicio de publicaciones de la Universidad Autónoma de Madrid.

morfosintácticos y búsquedas de palabras o estructuras en sus contextos de aparición". (Campillos Llanos, Gozalo Gómez y Moreno Sandoval, 2007: 3).

Para conseguir este objetivo, los autores dotaron al corpus de una base teórica similar al de una gramática pedagógica y fueron construyendo un software de búsqueda que permitiera al docente seleccionar aquellos contenidos más interesantes para trabajar con sus alumnos. En general, pese a todos los datos y los enfoques que la aplicación para ELE pueda tener, los autores han optado, en principio, por usarla en el desarrollo de la compresión auditiva, como proceso muy necesario para una correcta adquisición de la L2.

C-ORAL-ROM-ELE se inició con la elección de los contenidos gramaticales, categorías léxicas, nociones y funciones comunicativas que interesaba trabajar. Después se pasó al estudio de los fragmentos y documentos del corpus que se iban a utilizar por medio de una metodología basada en dos procedimientos básicos: la extracción de concordancias de una estructura lingüística y la lectura y escucha global de los textos.

Una vez estudiados todos los textos orales, se fueron clasificando según los niveles de dificultad basándose en el MCER y atendiendo, sobre todo, a las indicaciones que se hacían en él para la comprensión auditiva. De este análisis resultaron 92 documentos (de un total de 183 del corpus completo), cuyos niveles predominantes son el B2 (59 documentos) y el C1 (24 documentos).

Además, se ha pasado por una depuración y mejora de los documentos o fragmentos utilizados, eliminando errores ortográficos, y mejorando la puntuación. Se han añadido también nuevos datos a la cabecera de cada documento sobre contenidos gramaticales, nociones comunicativas, contenidos léxicos, dicción, velocidad de elocución, registro y nivel de dificultad. Finalmente, esto permitirá que la consulta del corpus pueda hacerse según criterios gramaticales, nocio-funcionales, de léxico o de cualquiera de los datos añadidos a la cabecera.

Cada categoría propuesta se ha asociado con una serie de fragmentos o documentos que fueran interesantes para su estudio hasta un máximo de 50. A través de ellos se puede contar con un material perfecto para el diseño de actividades con documentos de lengua oral, así como un estudio para distinguir sus rasgos característicos y el desarrollo

de habilidades de comprensión auditiva (perceptivas y cognitivas) y de estrategias (sociales, lingüísticas o de contenido) (Campillos Llanos, Gozalo Gómez y Moreno Sandoval, 2007).

La segunda fase del proyecto incluye la realización de un libro de actividades de comprensión auditiva para determinadas nociones comunicativas y aspectos gramaticales apoyándose en el uso de textos orales procedentes del corpus, previamente seleccionados y mejorados en la fase inicial. La finalidad de dicho libro es la de ofrecer un material complementario para los docentes y aprendientes de ELE, que podrán con él ampliar su trabajo en comprensión auditiva, principalmente, y de forma indirecta, ciertos contenidos gramaticales o léxicos.

Como sabemos, C-ORAL-ROM es un corpus oral de hablantes nativos, por lo que muchos de sus textos tienen una cierta complejidad. Por lo tanto, este trabajo se destina, en principio, a aprendientes de los niveles intermedios, avanzado y perfeccionamiento (que corresponden a los niveles habituales de B1-B2 y C1-C2 del MCER).

En total, se ofrecen más de 6.000 ejemplos presentes en más de 3.500 sonidos y se incluye una herramienta hipertextual que permite su consulta de forma rápida y sencilla y la escucha de todos los contenidos lingüísticos.

En la actualidad, el Laboratorio de Lingüística Informática de la UAM también trabaja, como hemos señalado anteriormente, en la elaboración de distintos corpus orales de aprendientes: un corpus de hablantes de ELE, tesis doctoral de L. Campillos, que recoge entrevistas a hablantes de distintas lenguas maternas de varios niveles del MCER, y otro de aprendientes de Japonés, realizado por E. Takamori.

6. Conclusiones

Sería muy deseable que el uso de corpus se generalizase en el ámbito de la enseñanza, no sólo porque conlleva numerosas beneficios para el proceso de aprendizaje, sino porque sería coherente con una sociedad que demanda cada día con más insistencia el aprovechamiento de unas competencias digitales diversas.

En el presente capítulo, hemos tratado de definir la relación existente entre los corpus y la enseñanza, desde un punto de vista general

y posteriormente, centrándolo en la enseñanza de lenguas. También, hemos presentado sus tres formas de uso principales (directo, indirecto y compilación de corpus pedagógicos), haciendo una descripción de cada una de ellas. Hemos dado cuenta de las ventajas, efectos y críticas más frecuentes a la utilización de corpus, y hemos comentado asimismo aquellos proyectos más sobresalientes en este campo en España y Francia.

En nuestra opinión, resulta necesario que los docentes comprendan que los corpus son un aliado excelente para el desempeño de muchas de las actividades pedagógicas que se llevan a cabo, porque además de ser una fuente inagotable de recursos, muestras de lengua o contextos, su uso conlleva un cambio en la metodología de enseñanza que comporta grandes beneficios para el aprendizaje, pues permite desarrollar competencias muy necesarias como el espíritu crítico o la conciencia lingüística. Con los corpus, el aprendiente pasa a tener un rol más activo en su aprendizaje, lo que conduce, sin duda, una adquisición más consistente de la lengua meta.

Los aspectos positivos del uso de los corpus en la enseñanza de las lenguas, no pueden hacernos olvidar ciertos problemas relacionados con su uso. Su forma y contenido es complejo y en muchos casos necesita de adaptación para que el aprendiente pueda realizar tareas provechosas. La riqueza de los contenidos, o por el contrario, la ausencia de fenómenos muy presentes en la lengua de uso habitual, pueden conducir a interpretaciones y generalizaciones erróneas, e incluso, a conformar una variedad de lengua de referencia demasiado pobre (si el corpus no está equilibrado ni es suficientemente representativo).

Sin embargo, aún siendo conscientes de los problemas que conllevan algunos corpus, podemos soslayar parte de estos aspectos negativos preparando adecuadamente los corpus y las actividades o tareas relacionadas con ellos, y dotando a los aprendientes del corpus más apropiado para su nivel. Pre-editar un corpus para adaptarlo a nuestros alumnos aumenta las posibilidades de éxito de la actividad.

De hecho, el uso de corpus está mucho más integrado en nuestras actividades diarias de lo que nos creemos, ya que, como bien señala McCarthy (2008), somos usuarios de corpus simplemente por utilizar Internet. La Web no es nada más y nada menos que un corpus inmenso, en el que establecemos búsquedas a través de un sistema que nos

proporciona un número determinado de concordancias, que, posteriormente, valoramos, analizamos, con espíritu crítico.

La relación entre corpus y enseñanza es aún algo inmadura en lenguas como el Francés y el Español, ya que existen trabajos muy aislados, poco representativos y, a menudo, demasiado modestos y escasamente difundidos. Sin embargo, no podemos dejar de valorar aquellos proyectos ejemplares cuyos autores, poco a poco, intentan difundir las ventajas de uso de los corpus y animar a los docentes a desarrollar nuevas aplicaciones o proyectos. En las lenguas románicas no se ha alcanzado al nivel de difusión del Inglés, lengua emblemática en el ámbito de la Lingüística de Corpus, pero existen ciertos proyectos de alto nivel para el español como C-ORAL-ROM-ELE o el corpus CEDEL 2 desarrollados en la UAM, y para el francés, con los corpus de aprendientes como FRIDA, los estudios de Boulton o la plataforma FLEURON.

Todos estos productos permiten ensayar una nueva metodología docente, ya que ofrecen numerosas posibilidades de explotación y la posibilidad de profundizar en mejoras pedagógicas en el ámbito de la adquisición de lenguas, algo muy oportuno en un momento en el que no se consiguen fácilmente los objetivos generales marcados a nivel europeo para las clases de lengua. Este tipo de actividades constituyen un complemento muy esperanzador, sobre todo porque inciden en aspectos de la lengua que normalmente no pueden ser explotados a través de actividades más convencionales, y resultan muy necesarios para adquirir una competencia efectiva en todos los contextos de uso.

Sin duda alguna, la compilación de corpus pedagógicos es una de las herramientas más efectivas que ha generado la aplicación de los corpus a la enseñanza. Un ejemplo claro lo constituyen los corpus de aprendientes, que inspiran muchos estudios, logrando desentrañar el camino que sigue el aprendiente a lo largo del proceso de adquisición, las características de su lengua intermedia o interlengua, y muchas de las dificultades a las que se enfrenta su aprendizaje. Toda esa información es de un gran valor para ayudar a los docentes a mejorar su práctica diaria y contribuir al diseño de nuevos programas de enseñanza de lenguas, mucho más efectivos y centrados en las necesidades y especificidades del aprendiente, como preconizan la mayoría de los currículos de enseñanza actuales.

4. EL APRENDIENTE

Ser un estudiante de idiomas puede consistir para la mayoría de la gente básicamente en haber realizado el acto de matricularse en un curso, pero los expertos no lo entienden ni mucho menos de una manera tan simple. Desde Piaget y Vygotski, el acto de aprender otra lengua implica un complicado proceso de construcción del conocimiento; la neurología, por ejemplo, nos enseña que cada individuo adopta estrategias de aprendizaje que le son específicas, y los psicólogos, por su parte, anteponen la motivación como factor esencial, pues de ella dependen tanto las actitudes como las representaciones. Excede el propósito del presente capítulo ofrecer una panorámica detallada de la investigación sobre los procesos de aprendizaje, pues solo pretendemos aquí centrarnos en el papel que desempeña el aprendiente en su propio proceso formativo.

1. Didáctica versus pedagogía

Aunque los términos *didáctica* y *pedagogía* suelen utilizarse indistintamente (por ejemplo: pedagogía de las lenguas/didáctica de las lenguas), nos parece que un uso riguroso obliga a tener en cuenta que para la didáctica, lo determinante para enfocar los procesos de adquisición y estructuración del conocimiento es la especificidad de la materia. La pedagogía, por el contrario, centra más su reflexión en la eficacia de la relación entre el profesor y el alumno, o entre los propios alumnos.

De lo que se deduce que *la didáctica* es del ámbito del estudio de las materias escolares o académicas específicas (dividiéndose en *enfoques* o *metodologías de enseñanza o de enseñanza/aprendizaje*), mientras que *la pedagogía* es más generalista. Así, nos parecería apropiado, por ejemplo, expresarse en términos de *teorías pedagógicas aplicadas a la didáctica de la lengua*.

Por nuestra parte, extendemos la famosa afirmación de Durkheim según la cual "la pedagogía es una teoría práctica" (Durkheim, 1922) al campo de la didáctica, y no descubriremos nada nuevo, apuntando

además que tanto la una como la otra son susceptibles de estudio científico. La presente investigación se encuadra, de hecho, en el campo específico de la didáctica de la enseñanza-aprendizaje de las lenguas no maternas.

1.1 El concepto de aprendiente

El anterior preámbulo sirve de base para comentar lo que tal vez resulta una obviedad, que es que los conceptos de *alumno* y de *estudiante* pertenecen al ámbito de las prácticas educativas, mientras que los de *sujeto de aprendizaje* y *aprendiente* son más propios de la jerga científica.

La idea de *alumno* tiende a asociarse a la dinámica pedagógica tradicional basada en la transmisión lineal y mecánica de conocimientos, mientras que la de *estudiante* remite a un agente activo, lo cual engloba a su vez las nociones de *procesos cognitivos*, *autoaprendizaje* y *afectividad*.

En cuanto al binomio *sujeto de aprendizaje/aprendiente*, aunque ambos términos siguen utilizándose alternativamente, a veces como sinónimos y otras con propósitos de distinción teórica específica, en general nos parece que su uso puede estar muy influenciado por el debate filosófico y sociológico en torno *al papel del sujeto* en la postmodernidad (¿hasta qué punto el sujeto es consciente, racional, autónomo, coherente, estable….?). De confirmarse, esta teoría podría explicar la predilección por el uso menos comprometido de *aprendiente* en la literatura de la didáctica de las lenguas de las últimas décadas. *Aprendiente* es, en todo caso, el término que se utiliza en el marco del presente trabajo para referirse al individuo que aprende una lengua que no es su lengua materna.

2. La situación de aprendizaje

La definición de la *situación de aprendizaje* no se agota con una presentación detallada de la ecuación alumno-profesor-conocimiento, a la que se le acoplarían unas concepciones particulares de los procesos de aprender y enseñar. En realidad, la última década se caracteriza por la aparición de una gran diversidad de enfoques metodológicos basados en

concepciones diferentes del proceso de aprendizaje, por lo que nos ha parecido que la mejor forma de sacar partido de tan variadas propuestas sería fijarnos, en primer lugar, en los marcos conceptuales que permiten analizarlas.

Entre estas teorías sobre los marcos conceptuales de la situación de aprendizaje, resulta muy habitual la utilización de la noción de ***triángulo pedagógico,*** que parte de una propuesta de Houssaye, (Houssaye, 1988) según la cual todo acto pedagógico se apoya en unas relaciones que es posible inscribir en los lados de un triángulo:

- **La relación entre el profesor y el saber**, que incluye conocer, saber hacer, saber estar, saber actuar, saber transmitir…, y que es la que permite *enseñar*;
- **La relación entre el profesor y el alumno** que permite *formar*;
- **La relación entre el alumno y el saber** que permite *aprender*.

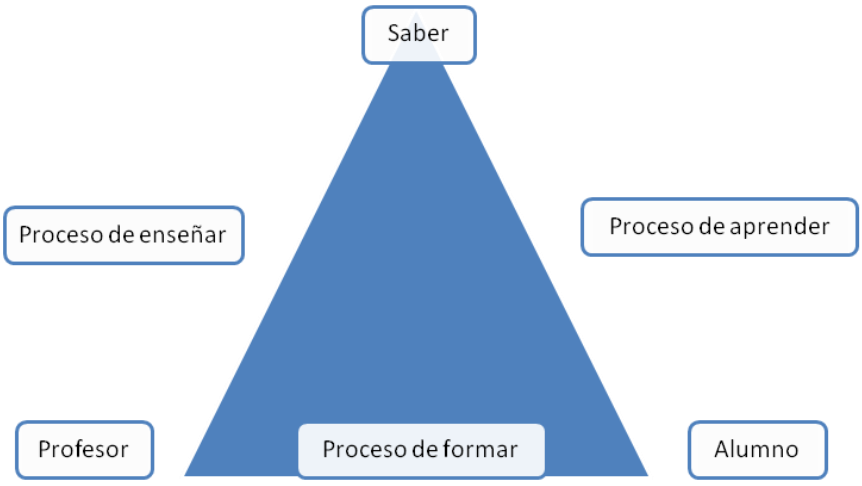

Gráfico 1: Representación del acto pedagógico según las teorías de Houssaye. Fuente: Elaboración propia

La tesis de Houssaye (2000) es que todas las situaciones pedagógicas se sustentan siempre de manera privilegiada en dos de los

lados del triángulo (sujetos activos) en detrimento del tercero (sujeto pasivo). Es lo que el autor llama el principio del *tercero excluido*. Por ejemplo:

- **la enseñanza tradicional** considera que el saber es la razón de ser del sistema educativo, por lo que se apoya fundamentalmente en ese saber y en el profesor que lo detenta, desatendiendo a los alumnos que se limitan a tomar notas, memorizar o utilizar las fuentes que les indica el profesor.
- **en los métodos no directivos** (pedagogía de Dewey, Freinet, Rogers…), por el contrario se da prioridad a la relación profesor-alumnos en la que el profesor asume un papel de mediador. En cuanto al saber, la programación de la materia puede resultar incluso inexistente ya que no se encuentra en el inicio del proceso de aprendizaje, sino que es la consecuencia de la actividad pedagógica.
- **en los enfoques basados en las TIC o en el *eLearning*,** es el profesor quien prácticamente desaparece del escenario.

El modelo de Houssaye ha sido criticado, por ejemplo, por Marguerite Altet que lo considera demasiado rígido y propone como alternativa un *modelo sistémico* (Altet, 1997), el cual tiene más en cuenta la complejidad que caracteriza al proceso de enseñanza-aprendizaje. Según lo expresa la autora, el modelo sistémico "pone el acento sobre la dinámica de la regulación pedagógica que tiene que ver con el flujo, la energía y el tiempo, más que sobre el equilibrio entre los tres polos" (Altet, 1997).

Por su parte, Legendre (1988) presenta el modelo conocido como SOMA, que se centra en las relaciones pedagógicas que se encuentran en el centro de la situación de aprendizaje, de enseñanza y de didáctica, en una situación pedagógica determinada:

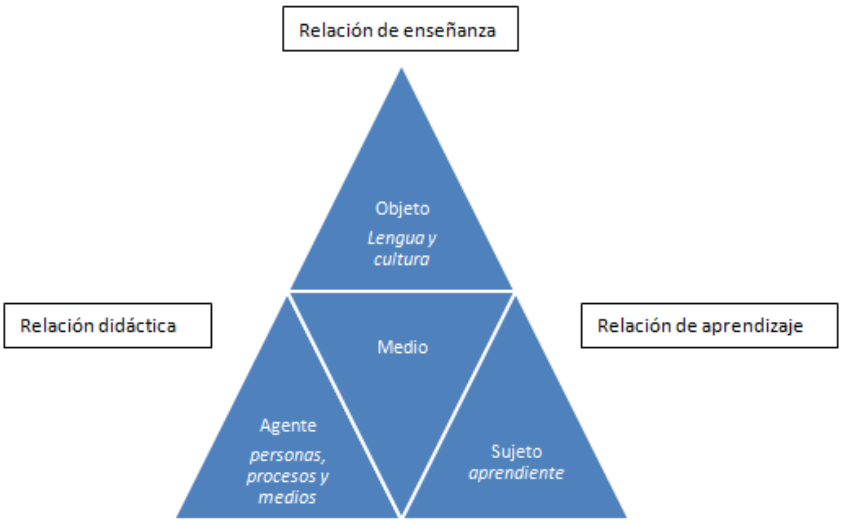

Gráfico 2: Modelo SOMA de Legendre. (Fuente: Elaboración propia)

En este modelo, encontramos:

- **El sujeto** (S), que es la persona en situación de aprender;
- **El objeto** (O), que es el objetivo del aprendizaje;
- **El medio** (M), que se refiere al entorno educativo humano (profesor, tutor, orientador…), las operaciones (inscripción, evaluación…) y los medios materiales (locales, muebles, medios económicos…)
- **El agente** (A), que remite a los recursos que asisten al aprendiente (profesor y compañeros), los medios pedagógicos (libros, ordenadores…) y los procesos (actividades individuales o colectivas, clases magistrales…).

Será Germain (1993, p. 10-12) quien aplique el modelo SOMA a la enseñanza de las lenguas ofreciendo una estructura analítica de los enfoques situacionales y comunicativos, que contribuye a reducir la brecha entre las teorías y las prácticas, al permitir analizar en profundidad los elementos presentes en la situación de aprendizaje.

3. Los estilos de aprendizaje

¿Por qué algunos individuos prefieren aprender una lengua a través de los textos, otros necesitan toda clase de explicaciones teóricas para poder progresar, mientras otros aprenden más rápidamente a través de la práctica y la introspección?

En Estados Unidos se ha desarrollado desde el final de los años 80 una investigación profusa sobre los estilos cognitivos que, a su vez, ha dado lugar a propuestas muy variadas de clasificación de los *estilos de aprendizaje*. El objetivo de los citados trabajos es contribuir a la adaptación de las metodologías de enseñanza a las características de los aprendientes, con la intención, lógicamente, de optimizar los resultados de aprendizaje. En los modelos de formación basados en el mejor conocimiento de los estilos de aprendizaje, junto a los aspectos cognitivos, se tiende a conceder una importancia cada vez mayor a los aspectos afectivos.

En general, se entiende por *estilo de aprendizaje* el modo individual espontáneo con el que un individuo asimila las informaciones y las trata en diferentes situaciones (experiencias) de aprendizaje de la lengua. En la literatura pedagógica de las lenguas, los estilos de aprendizaje suelen considerarse subcategorías de los *estilos cognitivos*, aunque lo cierto es que, dependiendo de los autores, *estilo de aprendizaje* y *estilo cognitivo* no siempre están bien diferenciados.

En realidad, varios autores destacan la dificultad de encontrar una definición consensuada del concepto de *estilo de aprendizaje* (Curry, 1990b; Riding y Rayner, 1998). Adrianne Bonham (Bonham, 1987) sostiene que hay una verdadera confusión al respecto. Dependiendo de la perspectiva de cada autor, el estilo de aprendizaje puede tener que ver con:

- Un conjunto de características personales

Para los autores que defienden este punto de vista, el estilo de aprendizaje se refiere a la modalidad personal, constante, distintiva y relativamente estable de comportarse desde el punto de vista psicológico, afectivo y fisiológico, en una situación de aprendizaje (Keefe, 1987: 36).

- La manera de abordar un conocimiento nuevo

Para Dunn y Dunn (Dunn y Dunn, 1993: 2) es el modo en que un aprendiente se concentra en una información nueva y difícil, la trata y la retiene. Para estos autores, el estilo de aprendizaje se basa en 18 elementos básicos, y es único para cada aprendiente, por lo que no permite tipificarlo. Dunn y Dunn defienden que rara vez dos personas aprenden exactamente de la misma manera.

En las dos definiciones anteriores, se pone el acento en el proceso de aprendizaje, más que sobre la competencia del individuo o el resultado del aprendizaje.

- Una personalidad determinada

Para Reinert (Reinert, 1976: 161), el estilo de aprendizaje es el modo en que una persona está programada para aprender más eficazmente, es decir que se refiere a su competencia para asimilar, retener y utilizar una información nueva.

- Predisposición para adoptar estrategias

El estilo de aprendizaje puede ser concebido como la disposición de algunos aprendientes para adoptar estrategias particulares al margen de las tareas concretas de aprendizaje que le son encomendadas (Kolb, 1984: 67; Das, 1988: 101; Schmeck, 1983: 233).

En estas dos últimas definiciones, se hace hincapié en las preferencias individuales, pero las cuatro definiciones contempladas hasta ahora tienen en común que se observa el modo en que la información es tratada por el aprendiente, pero no se entra a analizar la eficacia de una u otra modalidad.

Para Hunt (Hunt, 1979: 27) definir el estilo de aprendizaje de un aprendiente sirve para decir qué enfoque sería el más apropiado para él. En una definición más reciente, Riding y Rayner (1998) insistían en asociar el estilo de aprendizaje con un tipo de personalidad, y unos procesos prototípicos de tratar la información y actuar en contexto de aprendizaje. Thomas y Harri-Aunstein (1990) precisan que esas conductas habituales son repetitivas, automáticas e inconscientes.

Chevrier, Fortin y otros (Chévrier, Fortin et al., 2000: 10) lamentan que la noción de estilo de aprendizaje esté evolucionando actualmente hacia la dispersión, en lugar de hacerlo hacia la estructuración. Afirman que la mayoría de las definiciones parten de un aprendiente reactivo, sin un dominio real sobre el proceso de aprendizaje, esclavo de su propio estilo.

3.1 Estilos de aprendizaje según David Kolb

Se considera a David Kolb como uno de los teóricos que más ha contribuido al estudio de los estilos de aprendizaje. En 1984, Kolb publica su ensayo *Experiential Learning* en el que explica el principio a través del cual un individuo realiza su aprendizaje a través del descubrimiento y la experiencia.

Según sus observaciones, toda persona que está en proceso de aprendizaje pasa por cuatro fases:

1ª fase: experiencia concreta de una acción/idea

2ª fase: Observación atenta y reflexiva.

3ª fase: Conceptualización abstracta y teórica

4ª fase: Puesta en aplicación de la idea/acción en función de la experiencia inicial.

Partiendo de esa base común, Kolb demuestra que entre los adultos existen cuatro estilos de aprendizaje:

- El estilo acomodaticio

Al acomodaticio le gusta ejecutar tareas y tiende a aprender mediante el ensayo/error, más que mediante la lógica: se siente cómodo en los juegos de rol, en las actividades en grupo y en los debates y comentarios.

- El estilo divergente

El divergente tiene un sentido agudo de la observación. Examina los problemas bajo diferentes ángulos y destaca en las actividades de tormenta de ideas. Tiene mucha imaginación e intereses muy variados. Se interesa por las personas y es muy sensible a los sentimientos. Sus actividades favoritas serán aquellas que le permiten tener nuevas experiencias, intercambiar y discutir, tener clases particulares, observar y sacar conclusiones.

- El estilo asimilador

El asimilador destaca por su habilidad para organizar informaciones dispares. Se siente más atraído por la teoría que por la práctica. Le interesan las clases teóricas, los exámenes objetivos y las lecturas sobre diferentes temas teóricos.

- El estilo convergente

El convergente es el más hábil en llevar las teorías a la práctica. Prefiere los problemas que tienen una resolución única, mejor si son de tipo técnico. No se enreda con controversias personales o actividades grupales. Tiene predilección por el estudio y las actividades no dirigidos y los estudios de casos.

3.2 Características previas del aprendiente

En el apartado anterior, comentamos las teorías sobre los estilos de aprendizaje desde el punto de vista de la pedagogía general. Según la didáctica de las lenguas, las modalidades individuales de aprendizaje dependen principalmente de las siguientes características previas del aprendiente:

- **Conocimientos anteriores**

  El saber, saber hacer, saber ser y saber estar del aprendiente, es decir su capacidad previa para afrontar experiencias nuevas e

integrarlas a sus conocimientos anteriores, predisponen su aptitud a adquirir los nuevos conocimientos. Como es bien sabido, el aprendiente construye el sentido de los nuevos conocimientos, estableciendo lazos con los anteriores. Por eso, resulta importante que la estrategia de aprendizaje encuentre el modo de detectar los indicios de esa base previa, para cimentar sobre ella la ampliación del nuevo bagaje cognitivo.

- **La competencia de comunicación**

Tradicionalmente, era muy común organizar la estrategia de enseñanza partiendo de las llamadas *cuatro competencias*: comprensión oral, comprensión escrita, expresión oral y expresión escrita. Lo más dañino de este planteamiento fue que con frecuencia propiciaba la privación de un aprendizaje integrado, cediendo el profesor a la comodidad de desarrollar prácticas aisladas específicas para cada una de las citadas competencias, con el objetivo ilusorio de que la suma de todas ellas daría como resultado una competencia de comunicación. Hoy se habla incluso de "la muerte anunciada de las cuatro competencias" (Rosen, 2005).

Cuestionando la dicotomía entre *competencia* y *realización* de Chomsky por reducirse en exceso al campo gramatical, Hymes (1972) se refiere a *las variedades y estilos en la manera de hablar constitutivamente heterogéneos* de las comunidades lingüísticas reales. El concepto de *competencia de comunicación* de Hymes creado en su origen para su uso en el campo de la etnografía de la comunicación, se introdujo de una manera muy fértil en el terreno de la didáctica de las lenguas. Así, ya en 1978, Widdowson afirmaba que el objetivo final en materia de aprendizaje de las lenguas era la competencia de comunicación, a la que definía como la habilidad para interpretar de manera implícita (actividad mental subyacente a la capacidad de leer, escribir decir y escuchar) o explícita (actividad verbal) al conocimiento o la intuición especial para captar las reglas fonológicas, morfológicas, sintácticas y léxicas que rigen el uso de la lengua objeto, lo que incluye su aplicación en contexto comunicativo (Doyle & Rutherford, 1984). De la competencia lingüística depende la progresión de las capacidades de comprensión y de producción lingüística.

- **La atención**

  Los propios estudiantes son conscientes de la importancia de su nivel de concentración y de las consecuencias de la distracción sobre el éxito del aprendizaje. Pero la atención es a su vez dependiente del tema o contenido de aprendizaje, así como del tipo de actividad pedagógica.

- **La dimensión afectiva**

  La dimensión afectiva engloba las actitudes, las emociones y la confianza en uno mismo, las creencias y la cultura personal. En ella destacan aspectos tan importantes como la autoestima, la inhibición, la capacidad para asumir riesgos, la ansiedad, la empatía, la aptitud o la extroversión (Brown, 2007). Es esencial también hablar de la motivación como uno de los factores internos que tienen una mayor incidencia en el aprendiente (como bien señalan los estudios de Dörnyei). Por su parte, es importante que el profesor explique en lo posible las estrategias de comprensión y producción, es decir las diferentes maneras de afrontar las dificultades, y también que se muestre abierto a responder a todo suerte de preguntas y demandas de clarificación.

4. Estilos de aprendizaje y uso de corpus

El concepto de estilo de aprendizaje es, a la postre, inevitablemente confuso, si tenemos en cuenta que las teorías psicológicas son muy diversas y más aún sus interpretaciones por parte de sus analistas. Por otra parte, están relacionadas con una pluralidad de factores individuales (edad, sexo, personalidad, lengua materna, motivación, cultura…) que normalmente los investigadores tratan como un capítulo aparte, cuando no simplemente los ignoran. Así las cosas, se comprende las dudas sobre la conveniencia de utilizar los corpus en la enseñanza de las lenguas teniendo en cuenta los estilos de aprendizaje (Tyne, 2009, Flowerdew, 2008a y Boulton, 2008).

En cualquier caso, las investigaciones sobre la idoneidad del uso de corpus en relación con el estilo de aprendizaje parecen haberse limitado, hasta ahora, a la observación de la dimensión inductivo/deductivo. Por otra parte, la utilidad de tales indagaciones no es del todo evidente, si tenemos en cuenta que la tarea de consulta en corpus implica un proceso predominantemente inductivo, de lo que resulta lógico esperar que las mentes inductivas obtengan mejores resultados que las mentes deductivas. Tyne (2009) y Boulton (2010) destacan el interés de las conclusiones de un estudio de Lewis (2006). En primer lugar, porque este autor establece que en torno al 78% de los aprendientes tienen una mente deductiva, por lo que el uso de corpus no resultaría a priori el más adecuado para esa mayoría de personas. Sin embargo, parecen ser los deductivos quienes más manifiestan su interés por seguir trabajando con corpus (64% frente a 50% entre los inductivos), si bien desean hacerlo utilizando paralelamente las reglas tradicionales. Tal vez esa buena acogida del enfoque se explica por la apuesta metodológica hecha por Lewis: los enfoques alineados, es decir proponer actividades deductivas a aprendientes deductivos, e inductivas a aprendientes inductivos. Si se respetara este principio, los resultados serían óptimos.

Otras investigaciones intentan indagar la relación entre el uso de corpus en la clase de lengua y los estilos de aprendizaje (Flowerdew, 2008b, Turnbull y Burston, 1998, Felder y Spuling, 2005, Litzinger, 2007), pero se trata de estudios de casos con resultados difícilmente extrapolables además de previsibles. Como todo lo que tiene que ver con la investigación sobre el uso de corpus en la enseñanza de las lenguas, nos encontramos en este caso con una ausencia de resultados positivos. Pese a ello, hay que destacar que algunas tendencias parecen apuntar en una pluralidad de estudios: los perfiles más visuales (quienes realizan lecturas verticales de las concordancias alineadas) resultan sentirse más atraídos por este tipo de enfoque, mientras las mentes activas (propensos al manejo de datos) y las secuenciales (los lógicos que tienden a abordar los datos paso a paso) parecen ser las que sacan del uso de los corpus un mayor provecho que otros. Los perfiles menos idóneos serían los reflexivos, los verbales, quienes gustan de trabajar directamente con la lengua, y los globales, propensos a indagar ellos mismos las regularidades.

5. Conclusiones

A lo largo de este capítulo hemos realizado un breve resumen de lo que consideramos por aprendiente, y cómo este influye en la situación de aprendizaje. Además, hemos expuesto algunas de las teorías sobre su inserción y su papel en la situación de aprendizaje, destacando las teorías de Houssaye y Legendre.

Así, hemos hecho especial hincapié en el concepto de estilo de aprendizaje y en las distintas definiciones que se le han dado. Como sabemos, muchos son los autores que han intentado definirlo, no quedando muy clara la distinción entre estilo cognitivo y estilo de aprendizaje. De todas ellas destacamos la categorización de Kolb, que percibe al sujeto en proceso de aprendizaje como fruto del descubrimiento y la experiencia, y que habla de cuatro estilos de aprendizaje concretos para adultos: acomodaticio, divergente, asimilador y convergente.

Posteriormente, hemos señalado qué características del aprendiente se interrelacionan con los estilos de aprendizaje, destacando algunas como los conocimientos anteriores, los factores relacionados con la personalidad, o dimensión afectiva, la atención y la competencia de comunicación.

Finalmente, hemos resaltado cómo puede influir el uso de corpus según los distintos estilos de aprendizaje, y si el uso de estos es positivo para el aprendiente. Así, es importante mencionar que no existen numerosos estudios que valoren la relación entre ambos, como ocurre, por otra parte, con muchas otras aplicaciones y enfoques que van surgiendo de la incorporación de las TIC en la educación. Poner en marcha estudios de caso que lo valoren es generalmente muy complicado, ya que implica el control sobre un gran número de variables (difícilmente controlables cuando hablamos de procesos internos del aprendiente) y la necesidad de un entorno de instrucción o de aprendizaje muy concreto, con un número elevado de participantes. Además, aquellos que se han realizado hasta el momento no obtienen resultados estadísticamente concluyentes al respecto. Se cree en los beneficios del uso de corpus, sobre todo por su incidencia en la motivación del aprendiente ante una tarea distinta y que lo implica

directamente en el aprendizaje, pero todavía se apoya en análisis demasiado subjetivos.

5. CORPUS DE APRENDIENTES

1. Introducción

Un corpus puede recoger distintas variedades de lengua. Encontramos así corpus de hablantes nativos, corpus de lenguas de especialidad, -en el campo del derecho o de los negocios-, o corpus conformados por la alocución de hablantes que utilizan una lengua que no es su lengua materna. Estos corpus de hablantes no nativos, se crean normalmente a partir de las producciones orales de hablantes en proceso de aprendizaje de una lengua. Si dicho proceso tiene lugar en contexto educativo, a estos hablantes se les llama *aprendientes* y los corpus, a su vez, pasan a denominarse *corpus de aprendientes*. Con este uso del término, nos alineamos con trabajos desarrollados en otras lenguas que utilizan la expresión *learner corpora* en inglés o *corpus d'apprenants* en francés.

Por sorprendente que parezca, los corpus de aprendientes no son una idea reciente. Siempre ha habido profesores que han guardado las producciones escritas de sus alumnos, o que han grabado actividades de producción oral con el propósito de poner una nota que atienda a la progresión del alumno durante el curso, con la intención de encontrar fórmulas para mejorar la metodología de enseñanza o, simplemente, para identificar dudas, fallos u otros problemas de los alumnos a la hora de utilizar la lengua meta. De esta forma, hacia los años 80, con el desarrollo de las tecnologías y el surgimiento de la Lingüística de Corpus y de las técnicas para la implementación de los mismos, el uso de corpus en la enseñanza de lenguas resurge con las características y formato digitales.

Muchos de los corpus actuales también son realizados por los docentes con sus propios alumnos, -con fines de mejora del enfoque pedagógico, de las actividades o de la dinámica de las clases-, o con propósitos de investigación empírica (algunos se les conoce como *research action corpora)*. Además, una mayoría de estos corpus se dedican al análisis contrastivo pero, sobre todo, al análisis sistemático de errores. Lo que se pretende con ello es tener una visión más fidedigna de la interlengua de los alumnos.

Nuestro estudio, como recordaremos, presenta un corpus de aprendientes hispanohablantes de FLE. En el presente capítulo explicaremos con detalle en qué consiste este tipo de corpus, qué lo caracteriza y las aplicaciones más frecuentes que se derivan de ellos. Además, realizaremos una descripción de los corpus de aprendientes más importantes que se conocen hoy en día, atendiendo, como es lógico, especialmente al francés.

2. Definición

Un corpus de aprendientes es una colección de textos digitales que contiene producciones orales o escritas de hablantes que están aprendiendo una lengua que no es su lengua materna, por lo que es conocida habitualmente por *segunda lengua* (L2) o *lengua extranjera*. Estos bancos de textos son recogidos conforme a unos criterios de diseño determinados en función de su finalidad. Granger (2002) los define de la siguiente manera:

> *Computer learner corpora are electronic collections of authentic FL/SL data assembled according to an explicit design criteria for a particular SLA/FLT purpose. They are encoded in a standardized and homogeneous way and documented as to their origin and provenance* (Granger et al, 2002: 7).

El proceso de aprendizaje de la lengua que se desarrolla en la mente del aprendiente no es observable de manera directa, pero sí a través de análisis psicolingüísticos y neurolingüísticos. Los resultados de este tipo de estudios, pueden verse reforzados por el análisis de los datos fehacientes de producción, es decir, de los textos espontáneos orales o escritos producidos por los aprendientes. Se trata de realizar estudios longitudinales, principalmente, sobre el orden de adquisición de estructuras y sobre el uso de morfemas, elementos que preocupan especialmente a los investigadores del proceso de adquisición de L2.

Lo que conseguimos con ello es que la información procesada a partir de la interlingua nos permita poner a prueba de forma empírica nuestras hipótesis sobre los procesos de aprendizaje, o descubrir, en su caso, dificultades subyacentes a los procedimientos de adquisición que

no serían detectables por otros medios. Dicho conocimiento es la base para poder inferir de forma positiva en los procesos psicológicos, intelectuales o culturales que moviliza el aprendiente.

Las dos áreas de investigación en las que los corpus de aprendientes han sido más utilizados son los estudios sobre los procesos de adquisición de una L2 y la didáctica o metodología de enseñanza de L2. Sylviane Granger considera incluso que los corpus de aprendientes son una especie de puente entre ambas disciplinas (Cf. Aijmer, 2009: 13).

Al igual que otros tipos de corpus, los corpus de aprendientes han reportado grandes beneficios a la metodología científica. Proporcionando a los investigadores un gran volumen de datos provenientes de hablantes o informantes diversos, facilitan una sólida base empírica para la validación de hipótesis o generalizaciones relacionadas con los procesos de adquisición. Sin los citados corpus, no hubiese sido posible conocer mucho de lo que hoy se sabe sobre las fases del desarrollo de una nueva lengua.

Pese a ello, Granger (Granger, 2009) recalca el hecho de que todavía no se encuentre disponible el número significativo de aplicaciones pedagógicas concretas basadas que cabía esperar. La autora destaca tres razones para explicar la escasa deriva práctica de los corpus de aprendientes:

1. El campo de la adquisición de L2 ha estado dominado durante mucho tiempo por la corriente generativista, por tanto, por el convencimiento de la existencia de una gramática universal como factor determinante del desarrollo del lenguaje, por lo que los paradigmas basados en el análisis de datos objetivos no han gozado de demasiado predicamento.
2. El diseño de aplicaciones o herramientas basadas en corpus de aprendientes es, indudablemente, muy específica y compleja, y requiere, por otra parte, un gran esfuerzo analítico previo.
3. Dicha investigación debe hacerse bajo la responsabilidad de una persona con un perfil competencial complejo, pues se espera de ella que sea una conocedora cualificada de:

- *la Lingüística de Corpus*: para estar en condiciones de desarrollar las tareas de recolección y diseño;
- *teorías lingüísticas generales*: para realizar adecuadamente el análisis de datos;
- *los estudios sobre adquisición de una L2*: para estar en condiciones de observar e interpretar los procesos de generación de una interlengua;
- *la didáctica de las lenguas*: en muchos casos, se trata del ámbito más natural de aplicación.

Granger muestra de forma muy gráfica el citado perfil en la siguiente figura:

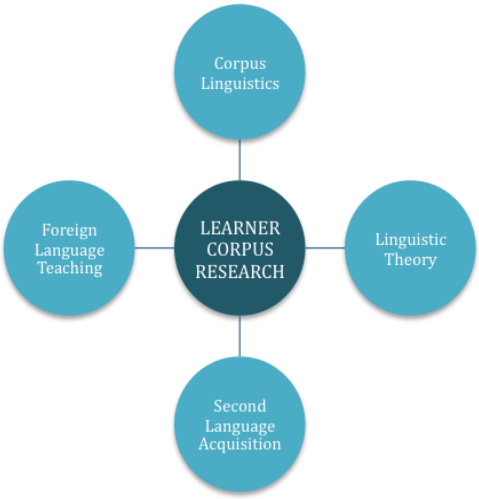

Gráfico 3: Core components of learner corpus research (Sylviane Granger). Fuente: Aijmer, 2009:15.

Aunque el planteamiento de Granger es relativamente reciente (2009), debemos decir que la tendencia actual es hacia la simplificación, ya que los corpus de aprendientes han conseguido que la colaboración entre expertos de esas distintas disciplinas sea cada vez más frecuente, redundando en un número creciente de proyectos, teorías e investigaciones.

3. Características básicas

Un corpus de aprendientes, como queda explicado, suele incluirse entre los considerados corpus específicos, que se ocupan de variedades de lengua concretas, en un contexto determinado o para unos fines particulares. Aunque podemos encontrar corpus de aprendientes de distintas disciplinas (como la biología, la economía o el derecho, por ejemplo), lo más frecuente es que lo relacionemos con los aprendientes de una L2[77]. De modo que, los aprendientes de una segunda lengua son un tipo de hablante específico, porque, aunque el objetivo último sea que su competencia sea asimilable a la de un nativo, su registro siempre será el específico de un aprendiente no nativo.

Además de específicos, los corpus de aprendientes suelen considerarse corpus *ad-hoc* o (*monitor corpus)*, porque nacen con un fin determinado, en un ámbito concreto, y su tamaño es variable, -relativamente reducido-, pues va ampliándose en función de las necesidades de la investigación en curso. La mayoría de los corpus actuales de aprendientes del ámbito académico oscilan alrededor de las 30.000 palabras, y reflejan la interlengua de un aprendiente local, lo que lo hace aún más específico. La excepción la ponen los corpus de aprendientes comerciales, creados por editoriales, que alcanzan más de un millón de palabras. Siguiendo a Granger (2004) y Pravec (2002), dicha especificidad relativa al tamaño nos lleva a distinguir entre:

- **Corpus de aprendientes académicos**, realizados por investigadores en el ámbito de un proyecto o investigación concreta y por lo tanto, *ad hoc* o muy específicos. Suelen ser de tamaño reducido, muy heterogéneos, y es muy difícil acceder a ellos porque son de carácter privado, siendo propiedad del grupo que los creó (salvo contadas excepciones);

[77] En ocasiones, también podemos encontrar corpus de aprendientes de una L2 limitados a un ámbito muy específico de aplicación, como la escritura en contexto educativo superior, la traducción de textos científicos, etcétera.

- **Corpus de aprendientes comerciales o de referencia,** realizados mediante la financiación de grandes editoriales, aunque desarrollados generalmente por expertos en Lingüística de Corpus y adquisición de segundas lenguas (ASL). Suelen alcanzar más de un millón de palabras, son muy variados y normalmente son de libre acceso (aunque no de forma gratuita). Pueden servir para complementar los materiales de referencia de las editoriales, como diccionarios, gramáticas y libros de texto. Este modelo de corpus no es muy frecuente, y prácticamente sólo existe para el inglés. Los más representativos, y quizá los de mayor de tamaño, sean el *Cambridge Learner Corpus*[78] y el *Longman Learners' Corpus*[79].

Al igual que el resto de corpus existentes, los corpus de aprendientes pueden ser catalogados según diferentes criterios: orales u escritos, monolingües o multilingües, paralelos o comparables, etcétera. También resulta habitual distinguir distintas tipologías de corpus de aprendientes en función del nivel de dominio de la lengua, de la lengua materna, o de la tarea que llevan a cabo los aprendientes en las muestras que conforman el corpus (interacción oral, producción escrita, resumen, descripción de imágenes, examen, etcétera). Pero la mayoría de los corpus suelen ser de textos escritos de carácter monolingüe.

Diseñar y crear un corpus de referencia de aprendientes es, como ya apuntamos, realmente complejo. Los corpus de referencia contienen millones de palabras, un elevado número de informantes así como de muestras de textos orales o escritos. Lograr reunir un volumen importante de palabras procedentes de hablantes no nativos es tarea fácil y supone un enorme esfuerzo por parte del investigador. Primero, porque no hay demasiados aprendientes dispuestos a colaborar. Segundo, porque las producciones escritas u orales del hablante de L2 suelen ser de un tamaño menor y, por tanto, con un número de palabras menor que las de un hablante nativo, por lo que se precisan muchos más hablantes para alcanzar una cifra total de palabras similar a la que reúnen los corpus de referencia. Y tercero, porque para que sea representativo y

[78] Disponible en la propia editorial: http://www.cambridge.org/gb/elt/catalogue/subject/custom/item3646603/Cambridge-International-Corpus-Cambridge-Learner-Corpus/?site_locale=en_GB

[79] Este corpus ha sido utilizado, además, para completar el *Corpus Learner Dictionary*: http://www.pearsonlongman.com/dictionaries/corpus/learners.html

equilibrado, suelen buscarse informantes con unas características concretas (por ejemplo, de una determinada edad, con un nivel de lengua en un contexto educativo específico, con los mismos años de experiencia en el aprendizaje de la L2, con un bagaje en otras lenguas extranjeras similar…), por lo que reunir a una cohorte suficiente de informantes resulta considerablemente difícil[80].

Ahora bien, contar con un corpus de tamaño reducido no implica que este deje de ser válido o representativo. Por un lado, porque todo corpus sigue unos determinados parámetros que controlan su diseño e implementación de una forma muy estricta que está ligada a las necesidades de la investigación en la que se enmarca. Todas ello contribuye a aumentar la representatividad de las muestras de lengua del contexto de estudio. Por otro lado, el gran tamaño de un corpus tampoco constituye una garantía de su validez para el análisis. Puede darse la circunstancia de que un repositorio de millones de datos implique una mayor dispersión y produzca frecuencias menores de aspectos o patrones concretos, e incluso una ausencia de los mismos. Algunas investigaciones como la llevada a cabo por Sinclair (Gavioli and Aston, 2001), así lo confirman, al demostrar que incluso los grandes corpus de referencia como, en este caso, el *British National Corpus*, no contenía ciertas expresiones que él necesitaba encontrar para su análisis.

Por si fuera poco, a la complicación que supone controlar simultáneamente muchas variables para la elección de sujetos y la implementación del corpus, hay que sumarle que es mucho más costoso para el investigador. Este aspecto ha contribuido también sin duda a que el número de corpus de aprendientes sea menor y, por lo general, de pequeño tamaño. Especialmente significativo es el caso de los corpus orales de aprendientes. Al no existir herramientas informáticas capaces de recoger y transcribir los archivos orales, se ha de realizar prácticamente todo el trabajo de forma manual por parte de los investigadores responsables. El esfuerzo que supone realizar un corpus oral (de cualquier tipo) y la dificultad de encontrar aprendientes de unas determinadas características explica que se trate aún de un campo escasamente explorado, que cuenta con escasísimas experiencias.

La mayoría de los corpus están, además, centrados en un determinado aspecto del proceso de adquisición, es decir, que son

[80] Por ello, investigaciones de centros académicos o científicos que producen los corpus de mayor tamaño coinciden con proyectos empresariales.

corpus sometidos a estudios trasversales. Por el contrario, los corpus longitudinales que permiten observar el desarrollo de un mismo sujeto de estudio a lo largo de un periodo de tiempo concreto, al ser más difíciles de recolectar, son mucho menos numerosos, o bien, de tamaño muy reducido. Una gran parte de los corpus, como ya hemos mencionado, nacen en el marco de un determinado proyecto de investigación. Un estudio longitudinal supone tener acceso a los mismos sujetos durante periodos de tiempo prolongados, algo que no pueden permitirse muchos de los proyectos de investigación actuales, que tienen una duración reducida, y obligan a los investigadores a realizar la implementación del corpus en el menor tiempo posible, para poder implementar el resto del proyecto (generalmente basado en los análisis del mismo) durante el periodo de tiempo estipulado.

Se sabe que la utilidad de los corpus de aprendientes se centra en las posibilidades que ofrecen de una aplicación pedagógica posterior. Así, según la previsión de utilización futura, Granger (Granger, 2009: 20) distingue dos subtipos de corpus:

1. *Los Corpora for delayed pedagogical use* (DPU), o corpus para uso pedagógico posterior, son aquellos en los que los aprendientes que han producido los datos, no los usan directamente como materiales de aprendizaje, es decir, como enfoques directos o DDL. Suelen ser implementados por editoriales o por investigadores con la intención de realizar un análisis de la interlengua o para elaborar, por ejemplo, materiales y aplicaciones más específicas para aprendientes de características similares a los informantes del corpus.

2. Los *Corpora for inmediate pedagogical use (IPU)* o corpus para uso pedagógico inmediato, suelen ser implementados por los profesores o investigadores como una actividad integrada en sus clases, en la que los aprendientes son a la vez productores y usuarios de los datos del corpus. Este subtipo es bastante reciente y mucho menos frecuente que los DPU.

Si comparamos ambos subtipos, observamos que los DPU son mucho más amplios, y por tanto, más aptos a garantizar unas generalizaciones representativas del conjunto de los aprendientes. Los IUP suelen ser, por el contrario, reducidos y de carácter marcadamente específico o local, por lo que no son representativos *per se*. Eso no quiere decir que no sean válidos, ya que, como muchos autores señalan, son quizá los más provechosos para los docentes, ya que permiten conocer en profundidad el proceso de adquisición y favorecen el desarrollo de la conciencia lingüística de los aprendientes que reflexiona sobre la lengua meta a través de sus propias producciones.

Cualquiera de los dos tipos resulta aún más útil si después de su implementación se realiza un trabajo de anotación. Es decir, si se le añaden ciertas notas o informaciones relativas a aspectos lingüísticos, a cuestiones interpretativas, de orden semántico, etcétera.

Una de las anotaciones más frecuentes por su relevancia dentro del campo de la adquisición de lenguas es, como hemos dicho, el análisis de errores, que puede marcarse dentro de la propia transcripción. Sin embargo, la anotación posterior suele ser un proceso manual, ya que la mayoría de programas existentes no son del todo eficientes a la hora de analizar la interlengua, debido, en gran parte, a los errores ortográficos o cambios sintácticos que contiene y que producen variaciones de la norma de lengua con la que está implementada el programa. Esto ocurre frecuentemente con otro de los tipos de anotación más frecuente, el etiquetado de las partes de la oración (*part-of-speech o POS taggers*), o cualquier otro tipo de analizadores sintácticos y/o morfológicos. Las características inherentes a la interlengua siguen siendo un obstáculo para alcanzar análisis automáticos eficientes.

Un último aspecto que podemos reseñar de los corpus de aprendientes es su disponibilidad. Evidentemente los corpus llamados comerciales están a disposición de la comunidad científica a través de la venta de una licencia o de un CD-ROM, pudiéndose utilizar para fines pedagógicos y de investigación casi sin restricciones. Los corpus de aprendientes realizados en el ámbito académico, sin embargo, suelen ser restringidos y muy pocos de ellos están disponibles. Algunos como ICLE, poseen CD-ROM o licencias para su uso, pero la mayoría ni siquiera han sido difundidos, sirviendo sólo para la investigación para la que fueron implementados. Muy pocos son los que aparecen en línea de forma gratuita, facilitando su consulta a través de un registro, y si lo

están, no suele haber un acceso al corpus completo, sino sólo a partes de él.

Sin duda, los corpus de aprendientes poseen numerosos aspectos beneficiosos para el aprendizaje, pero necesitan de una mayor regulación, estandarización y difusión para que se conviertan en una práctica más extendida entre la comunidad científica y docente.

4. Corpus de aprendientes de segundas lenguas y lenguas extranjeras representativos

Los corpus de aprendientes, como ya hemos comentado, son más frecuentes en el campo de la investigación en ASL y la didáctica de L2. Es en este segundo ámbito donde son más numerosos, principalmente por su utilidad para caracterizar la interlengua.

Los grandes corpus comerciales que conocemos sólo existen para el inglés, y son, principalmente dos: el *Cambridge Learners' Corpus* (CLC) y el *Longman Learner Corpus* (LLC)[81]. Ambos están realizados por investigadores de prestigio en el campo de la adquisición de lenguas y la lingüística de Corpus como son J. Sinclair y D. Biber. Abarcan millones de palabras y posteriormente a su creación, han dado lugar a aplicaciones pedagógicas concretas, como diccionarios específicos para aprendientes o gramáticas más adaptadas a sus necesidades específicas.

Los llamados corpus de aprendientes académicos, siguiendo la estela marcada por los corpus generales o de referencia, fueron primero implementados para caracterizar la lengua inglesa. En general, actualmente son mucho más numerosos que los corpus comerciales, aunque de tamaño menor, y en ellos podemos encontrar diversidad de lenguas, tipologías y finalidades, si bien la mayoría están enfocados a realizar análisis de la interlengua (para la investigación en adquisición de nuevas lenguas) y a aplicaciones pedagógicas posteriores. Entre estos, podemos destacar los siguientes:

[81] En este capítulo sólo vamos a comentar los corpus de aprendientes más importantes del ámbito académico por su relación con el objetivo de nuestra tesis. Los corpus comerciales de aprendientes serán citados, pero no entraremos en una descripción detallada de ellos al no existir ninguno consagrado al Francés como Lengua Extranjera (FLE).

4.1 Corpus ICLE

El *International Corpus of Learner English* (ICLE)[82] es quizá el corpus de aprendientes más extenso de todos los conocidos hasta el momento, y pionero en el ámbito académico. Fue realizado en el seno de un proyecto europeo en el *Centre for English Corpus Linguistics* de la Universidad de Lovaina[83] y dirigido por Sylviane Granger hacia los años 90, e incluye dos millones de palabras procedentes de trabajos escritos de alumnos de nivel avanzado con distintas lenguas maternas (los primeros fueron de lengua materna alemana y francesa).

En concreto, recoge muestras de hablantes de once lenguas maternas diferentes, entre las que se encuentra el español, el francés, el alemán, y el sueco. Posteriormente se completó con un corpus oral, la *Louvain International Database of Spoken English Interlanguage* (LINDSEI), que recoge producciones orales de hablantes de inglés como lengua extranjera de nivel avanzado en situaciones comunicativas estandarizadas, es decir, que todas siguen un determinado patrón de entrevista para facilitar la comparación.

Los objetivos principales de ICLE eran encontrar muestras de errores de los aprendientes y hacer estudios contrastivos entre lenguas para determinar cuáles eran universales y cuáles eran específicos de una lengua (materna) concreta. Además, pretendían llevar a cabo análisis sobre los aspectos en los trabajos escritos de aprendientes que se desviaban de un uso nativo de la lengua. Es decir, mostrar aquellos aspectos que separan a un hablante de L2 muy competente de un nativo, porque poseen un cierto matiz de "*foreign-soundingness*" (Pravec, 2002: 83), como la fraseología, el excesivo o reducido uso de ciertas estructuras, marcadores discursivos, entre otros.

Uno de los aspectos más interesantes de la investigación de ICLE fue su sistema de análisis de errores. Con la intención de poder generar herramientas y metodologías propias para los corpus de aprendientes, en el seno del ICLE se intentó desarrollar una herramienta informática de

[82] http://www.uclouvain.be/en-cecl-icle.html

[83] Debemos destacar que el *Centre for English Corpus Linguistics* de la Universidad Católica de Lovaina es un centro pionero en la investigación de los corpus de aprendientes, que ha estado siempre en la vanguardia mundial en este ámbito. http://www.uclouvain.be/en-cecl.html>

marcado y de edición de errores, llamada *Error Editor*, que permitía a los investigadores anotar textos con errores y establecer búsquedas o listas de errores frecuentes. Sin embargo, este editor, si bien parece haber sido utilizado, no ha tenido reflejo en ningún programa, plataforma o herramienta disponible para la comunidad científica.

El ICLE ha sido considerado el corpus de aprendientes de referencia en esta disciplina debido a que su completo diseño e implementación le dota de una gran representatividad y, por tanto, nos propone un arquetipo de aprendiente que puede ser cualquier aprendiente de inglés L2 de nivel avanzado.

4.2 Corpus de aprendientes de francés como lengua extranjera (FLE)

En el ámbito del francés como lengua extranjera (FLE), no encontramos aún corpus de aprendientes de referencia o comerciales. Es posible que no hayan trascendido por no haber sido difundidos por sus autores, o porque realmente no han suscitado un interés suficiente para que se difundieran.

Sin embargo, sí que podemos hablar de ciertos corpus de aprendientes de FLE realizados entornos académicos y de investigación que han sido pioneros en el ámbito del francés. A modo de resumen, mostraremos en la siguiente tabla las características principales de los corpus de aprendientes existentes en la actualidad que incluyen al francés como L2:

Corpus	L2	L1	Tipología	Nivel	Palabras	Metodología	Autoría
Chy-FLE (*Cypriot Leaner Corpus of French*)	Francés	Griego	Escrito monoling.	Intermedio-Avanzado	Aprox. 250.000 En desarrollo	Producciones escritas argumentativas y descriptivas	Universidad de Poitiers y Universidad de Cyprus

COREIL Corpus	Francés Inglés	-	Oral bilingüe	-	-	Estudio de fonología frasal y entonación del inglés y del francés L2 a través de distintas tareas: descripción de imágenes, argumentación, lectura...	Universidad de Paris Diderot
Dire autrement	Francés	Inglés (en su mayoría)	Escrito	-	48.114	Textos narrativos, persuasivos e informativos	Dalhousie University (Canadá)
FRIDA (*French Interlanguage Database*)	Francés	Varias lenguas	Escrito		Aprox. 450.000	Del proyecto *Freetext*, pionero en corpus anotados con errores, para la realización de la herramienta ALAO con actividades y textos auténticos orientado a necesidades específicas de aprendientes	Centre for English Corpus Linguistics, Université Catholique de Louvain (Sylviane Granger)
FLLOC (*French Learner Language Oral Corpora*)	Francés	Varias lenguas	Oral	Varios niveles	-	Conjunto de siete corpus basados en distintas tareas y distintos niveles, incluso universitario. Audio y texto disponibles gratuitos para investigación, y posibilidad de realizar búsquedas *online*.	Universidad de Southampton /Newcastle. (Florence Myles y Rosamund Mitchell)
InterFra Corpus	Francés	Sueco	Oral	Varios niveles en universidad	-	A partir de entrevistas, y narraciones, videoclips e imágenes. Disponible el audio digital. Transcripciones en formato XML para comunidad científica.	Universidad de Estocolmo (Inge Bartning)
IPFC (*Interphonologie du Français Contemporain*)	Francés	Varias lenguas	Oral	Varios niveles	En desarrollo	Basado en lecturas en voz alta, repetición de palabras, entrevistas semidirigidas e interacciones entre dos aprendientes.	Univ. De Waseda, Rouen, Genève y Tokio.

LCF (*Learner Corpus French*)	Francés	Danés	Escrito	Intermedio y avanzado	490.000	Basado en textos argumentativos, periodísticos, cartas formales y resúmenes. En curso de realización.	K.U.Leuven Campus Kortrijk, UGent and Lessius
Lund CEFLE Corpus (*Corpus Écrit de Français Langue Étrangère*)	Francés	Sueco	Escrito	Varios niveles	100.000	Textos descriptivos y narrativos e historias basadas en imágenes. Sólo parte del corpus está disponible en línea.	Lund University (Suecia)
UWI (*Learner Corpus of University of West Indies)*	Francés	Inglés Jamaicano (criollo)	Oral	Varios niveles	-	Conversaciones procedentes de exámenes orales y de contextos informales	Universidad de New South Wales, Sydney, (Australia)
ESF Database (*European Science Foundation Second Language*)	Danés Inglés Francés Alemán Sueco	Varias lenguas	Oral	Varios niveles	-	Conversaciones espontáneas de cuarenta adultos inmigrantes trabajadores en Europa Occidental con hablantes nativos de sus países de acogida. Disponible gratuito en línea.	Max Planck Institut, Nijmegen, Netherlands
MeLLANGE (*Learner Translator Corpus*) (LTC)	Multiling.	Varias lenguas	Escrito	Traductores en prácticas	-	Textos legales, técnicos, administrativos y periodísticos. Posibilidad de búsquedas con herramienta en línea.	Universidad Paris Diderot
Padova Learner Corpus	Multiling. Francés Inglés Español	Italiano	Escrito	-	-	Actividades de CMC (Computer-Mediated Comunication) como contribuciones a debate, dossiers, resúmenes... Corpus longitudinal. En desarrollo	Universidad de Padua (Italia)

PAROLE (*Corpus PARallèle Oral en Langue Étrangère*)	Multiling. Inglés Francés Italiano (también datos de hablantes nativos)	Varias lenguas	Oral paralelo	Varios niveles	30.000	Audios con 68 hablantes adultos en contexto universitario. 5 tareas: resumen de video humorístico, descripción y resumen de una videosecuencia, resumen de una secuencia publicitaria, crear frases complejas a partir de imagen fija introduciendo conectores, y narración autobiográfica de un accidente.	Universidad de Savoie (Francia)
FLEURON (*Français Langue Étrangère Universitaire: Ressources et Outils Numériques*)	Francés	Varias lenguas y nativos de L1	Oral y escrito	Varios niveles	Aprox. 400.000 Previsión de 700.000	Parte de una base de recursos destinada a estudiantes extranjeros para antes o durante la estancia en una universidad francesa. 75 horas de grabación de entrevistas, conversaciones y en contexto informal. En desarrollo.	Universidad de Nancy (France)

Tabla 2: Resumen de los principales Corpus de Aprendientes que incluyen al Francés

En definitiva, los corpus de aprendientes del francés son relativamente escasos, al menos, en comparación con los existentes para el Inglés L2. La mayoría de ellos son corpus escritos, y están asociados a determinados proyectos de investigación, por lo que poseen características específicas que no favorecen su reutilización. Tampoco existe, como ya hemos comentado, un corpus de referencia para el francés, puesto que no observamos ningún ejemplo que alcance un tamaño similar al de los grandes corpus comerciales de aprendientes de *Cambridge* y *Longman* para el inglés. No obstante, y aunque no poseen millones de palabras, muchos investigadores consideran a los corpus *FLLOC* y *FRIDA* como los corpus de referencia de la disciplina y, sobre todo, como modelos a seguir a la hora de diseñar e implementar un nuevo corpus.

Finalmente, como habíamos señalado anteriormente, tampoco encontramos de momento indicios de creación de corpus de aprendientes del francés en nuestro país. Se han encontrado algunas referencias bibliográficas que aluden a investigaciones basadas en corpus del francés, pero no tenemos noticia de ningún corpus constituido y difundido como tal. En estos casos, se trata de corpus escritos específicos, como corpus de textos periodísticos o literarios, que abarcan la obra de un determinado autor. En alguna referencia, se hace alusión a un análisis de errores de producciones escritas de aprendientes hispanohablantes de FLE, pero, aparentemente no se ha constituido un corpus independiente para desarrollarlo[84].

5. Usos y aplicaciones frecuentes de los corpus de aprendientes

Los corpus de aprendientes han sido, hasta hace algunos años, un elemento poco utilizado en el ámbito de la adquisición de segundas lenguas pese a su enorme potencial.

[84] Cf. SANTOS MALDONADO, M.J. : *El error en las producciones escritas de francés lengua extranjera: análisis de interferencias léxicas y propuestas para su tratamiento didáctico.* Tesis doctoral, Universidad de Valladolid, Departamento de Didáctica de la Lengua y la Literatura, 2003. En este caso, utiliza datos extraídos de la plataforma de comunicación a través del ordenador*, Galanet.*

La complejidad que supone su creación e implementación (aún mayor en el caso de los corpus orales) lleva a muchos docentes e investigadores a desechar la idea de su uso. Sin embargo, en los últimos años, las mejoras tecnológicas han propiciado su aumento, aunque todavía no llegan al número y la frecuencia de uso en investigaciones o proyectos de los corpus considerados de hablantes nativos.

Una de las causas del desinterés por el análisis de corpus de aprendientes fue las críticas recibidas por una de las metodologías de análisis que más lo utilizaban, el Análisis Contrastivo (en adelante, AC). Y más concretamente, el Análisis Contrastivo de la Interlengua (popularizado en su versión inglesa: CIA, *Contrastive Interlanguage Analysis*).

El AC fue uno de los primeros análisis llevados a cabo en el campo de la adquisición de lenguas. Con él se pretendía comparar la lengua materna o L1 y la L2 del hablante para tratar de buscar las divergencias, los puntos en común y las posibles interferencias negativas que podía causar la L1 en el aprendizaje de la L2. El desarrollo de nuevas metodologías y enfoques mostró que la L2 no está realmente tan influida como se pensaba por la L1 (de hecho, el análisis de errores mostró que el origen de muchos de los errores en el aprendiente era de carácter intralingual y no interlingual, como veremos más adelante en este estudio), y que existía una cierta falacia al comparar la L1 y la L2 al tratarse de lenguas distintas, regidas por unas normas distintas.

Sin embargo, desde nuestro punto de vista, los AC entre la L1 y la L2 son los más provechosos para los docentes, primero, porque ofrecen una mayor facilidad para encontrar las muestras que comparar, y proporcionan numerosos estudios e hipótesis ya contrastadas. Pero además, porque aportan bastantes datos para identificar los aspectos léxicos, gramaticales y discursivos que diferencian a los aprendientes de la norma de la L2 (Cf. Aijmer, 2009: 19).

Pese a ello, el AC se fue abandonando, al no gozar de demasiada popularidad entre los investigadores. Pero se recuperó posteriormente, gracias a la emergencia del concepto de *interlengua*. Se demostró que no era necesario hacer una comparación entre la L1 y la L2 para definir la lengua del aprendiente, sino que bastaba con analizar la L2 a través de estudios trasversales o longitudinales. Podemos comparar sujetos con el mismo nivel de conocimiento de una L2 para hallar los procesos comunes, los obstáculos más representativos, establecer relaciones entre

hablantes de L2 con distinta lengua materna, para averiguar si siguen el mismo proceso de adquisición, etcétera. En todos estos análisis es donde los corpus de aprendientes encuentran una aplicación esencial.

Además, el análisis de la interlengua puede no centrarse exclusivamente en los errores. Se puede sin duda caracterizar la interlengua estableciendo un recuento de los ítems conseguidos para cada nivel de adquisición en función del currículo u objetivos pedagógicos propuestos. De hecho, para superar la llamada "falacia comparativa", se ha de basar no tanto en una perspectiva de la lengua meta, sino en un estudio profundo de la interlengua. Por ejemplo, teóricos como Ellis y Barkhuizen[85] señalan que es importante analizar tanto lo que los aprendientes hacen bien como lo que hacen mal.

Como cualquier otro tipo de corpus, se pueden utilizar de forma directa o indirecta, siendo la segunda la más habitual. Así, una gran parte de las aplicaciones que existen se relacionan con un fin pedagógico, sirviendo para la mejora de muchas aplicaciones y métodos de enseñanza.

Los corpus de aprendientes se consideran en general poco aptos para todo tipo de alumnos, por lo que se desecha su utilización en experiencias basadas en corpus o en DDL, primando a los corpus de referencia y los corpus de nativos en general, que, para la mayoría de docentes, promueven un modelo de lengua mejor y más adecuado, y el modelo meta deseable para todo aprendiente. En definitiva, los corpus de aprendientes han tenido mucha más proyección en el ámbito de la investigación, sobre todo, en el campo de la adquisición de segundas lenguas.

5.1 Análisis para el entorno educativo y científico

El estudio de corpus de aprendientes se enmarca dentro de las distintas corrientes de análisis de datos que son propias de la Lingüística Aplicada. Con las muestras procedentes de corpus de aprendientes, podemos realizar, por consiguiente, diferentes análisis como haríamos con cualquier otro de sus tipos:

[85] Cf. Ellis y Barkhuizen, 2005.

- Análisis cualitativos
- Análisis cuantitativos
- Análisis longitudinales
- Análisis de frecuencias de uso
- Análisis de errores (AE) o análisis computacional del error (CEA, o *Computer-aided Error Analysis*)
- Análisis contrastivos (AC)
- Etcétera

Sin embargo, podemos afirmar que han sido muy utilizados en análisis tanto contrastivos como de errores, en un intento de mostrar detalladamente los pasos que guían al aprendiente en la adquisición de una lengua.

Los corpus de aprendientes aparecen también en el análisis de necesidades pedagógicas, como bien señala Tono (Frankenberg et al., 2011), quien retoma la idea de Tudor de distinguir entre *Target Situation Analysis* (TDA) o Análisis de la Situación Meta y el *Present Situation Analysis* (PSA) o Análisis de la Situación Presente, siendo los corpus de aprendientes el objeto de análisis del segundo. En el PSA se muestra la competencia actual de los aprendientes, las diferencias o la distancia existente entre las metas a conseguir en L2 y su grado o nivel de lengua meta real. Con este tipo de análisis se consigue definir la interlengua y mostrar sus características representativas. Se asume que conociendo el nivel real y los problemas más frecuentes de los aprendientes, podremos definir mejor la enseñanza y que ésta sea más adecuada y productiva para ellos.

Y, por otro lado, este contexto de análisis PSA puede servir para asignar a nuestro alumno dentro del nivel de competencia que verdaderamente le corresponde como sugiere Osborne (Frankeberg et al., 2011: 181). Sabemos que el MCER, en lo que a expresión oral se refiere, se caracteriza por una cierta indefinición a la hora de distinguir los índices descriptores que corresponden a cada nivel. Muchas veces es difícil interpretar dichos indicadores y verlos reflejados en el aprendiente de forma clara. Algunos parecen muy similares, o no están bien definidos, por lo cual es costoso aplicarlos de forma sistemática.

Por tanto, el trabajo con corpus de aprendientes podría ayudarnos de forma objetiva a asignar el nivel correspondiente a cada alumno. Comparando matices cuantificables en corpus de hablantes nativos y de aprendientes, como es el caso de las autointerrupciones, las reformulaciones (generalmente medidas cada 100 palabras, tanto para las repeticiones como las reformulaciones sintácticas), las vacilaciones, uso de apoyos vocálicos, las pausas, es decir, todo lo que conocemos como disfluencias, y otros factores como la longitud de las *utterances*, la velocidad de producción (palabras o sílabas por minuto), el número de palabras producidas, etcétera, podemos establecer baremos que nos indiquen cuál es el porcentaje para cada nivel, compararlo con nuestros aprendientes y establecer así el nivel más aproximado de competencia lingüística que le corresponde.

También habríamos de sumarle otros índices cualitativos como el contenido informacional que refleja la producción, el contenido sintáctico (densidad, tipo de estructuras utilizadas, número de oraciones subordinadas o complejas, entre otros), la variedad léxica o la corrección (tasa de errores por cada 100 palabras).

Sin embargo, no deja de ser un aspecto controvertido, ya que necesitaríamos un número muy elevado de hablantes nativos y de aprendientes para establecer los perfiles básicos que nos servirán para comparar. Además, y como ya hemos indicado anteriormente, los programas que analizan los corpus orales aún no tienen la misma eficacia que para los corpus escritos, lo que podría, sin duda, falsear los datos obtenidos.

Por consiguiente, los corpus de aprendientes se han utilizado para estudios más básicos, que se alinean en dos sectores claros:

- En el análisis y descripción de la adquisición por parte del aprendiente de la L2, mostrando sus fases de desarrollo, el orden de adquisición de conocimientos y patrones morfosintácticas y otras características de la interlengua;
- El análisis de errores (AE) frecuentes de los aprendientes, de manera que se pueda mejorar la calidad de la enseñanza y adecuarla al perfil de los aprendientes.

En ambos casos, nos encontramos con un análisis ascendente (denominado en inglés *bottom-up* en la bibliografía específica), y que pretende, en resumen, *aprender del aprendiente*.

Como hemos citado anteriormente, podemos destacar al Análisis Contrastivo de la Interlengua (conocido por sus siglas en inglés, CIA) como uno de los que, hasta el momento, ha jugado un papel principal e imprescindible en la identificación de aspectos específicos de la L2. Suele concentrarse en dos tipos de comparación:

1. la de la lengua del aprendiente y las muestras de uno o varios corpus de referencia de hablantes nativos (L1→L2 o LN, Lengua nativa → IL, Interlengua);
2. la comparación entre distintos tipos o variedades de lengua de aprendientes (L2→L2).

Los análisis derivados de los corpus de aprendientes se han focalizado sobre todo en la descripción del exceso o poco uso de distintos aspectos morfosintácticos o léxicos en varios contextos, y casi siempre en contraposición al uso de los hablantes nativos. Las áreas más estudiadas suelen ser las colocaciones, las oraciones subordinadas, la interrogación, la negación y los conectores o conjunciones.

5.2 Aplicaciones para la implementación de materiales pedagógicos y herramientas de aprendizaje de lenguas

A menudo, los análisis realizados con los distintos corpus suelen generar conclusiones o hipótesis que se aprovechan después en libros y manuales de referencia, como diccionarios o gramáticas. Otro ámbito de aplicación pueden ser los manuales o libros de texto para la adquisición de lenguas, aunque no muy extendido por el momento, y por supuesto, con el auge de las TIC(E), en aplicaciones informáticas, donde tampoco se han obtenido los resultados esperados.

Los primeros en aprovechar sus ventajas han sido los diccionarios, quizá por la larga tradición del uso de corpus en lexicografía, que se han decantado por incluir notas específicas para aprendientes, resaltando fenómenos concretos que podían inducir al error, usos problemáticos o aspectos de lengua más difíciles de adquirir. El primero de ellos fue el *Longman Essential Activator* (1997), que se

complementa con los análisis provenientes del *Longman Learner Corpus,* y que crea notas de ayuda en algunas entradas para llamar la atención de los aprendientes hacia ciertos aspectos lingüísticos, semánticos o morfológicos que presentan una mayor dificultad. Después de este, encontramos muchos otros como el *Longman Dictionary of Common Errors*, o en diccionarios más generales como el *Longman Dictionary of Contemporary English* (2003) o el *Cambridge Advanced Learner's Dictionary* (2003), que se nutre a su vez del corpus *Cambridge Learners' Corpus*.

Su uso en gramáticas ha sido posterior, y ciertamente, no es aún demasiado popular. Casi todos los ejemplos que encontramos se relacionan con el *Cambridge Learners' Corpus*, que en los últimos años, ha inspirado algunas mejoras en las gramáticas realizadas por Carter y McCarthy para la propia editorial Cambridge.

Su uso en libros o manuales para la enseñanza de lenguas, como hemos comentado en el capítulo precedente, no deja de ser testimonial. Con la eclosión de la utilización de corpus en enseñanza de lenguas muchas editoriales barajaron la posibilidad de utilizar los corpus para establecer el contenido de los libros, pero por diversos motivos, no se llevó a la práctica. De hecho, de existir manuales basados en corpus, lo son basados en corpus de referencia, pero apenas en los análisis de corpus de aprendientes.

Tampoco es muy frecuente encontrar aplicaciones informáticas que incluyan corpus de aprendientes. La propia complejidad de la transcripción, en el caso de corpus orales, o de los textos del aprendiente que forman parte de los corpus escritos, hace que la mayoría de programas no reconozca o tenga grandes problemas para ejecutar las acciones requeridas. Los sistemas aún no están lo suficientemente entrenados como para desambiguar fácilmente las producciones escritas u orales de los aprendientes, ya que están concebidos a partir de una norma escrita muy cerrada y poco flexible, que encuentra numerosos errores o cambios en la forma que no es capaz de asimilar y que restan eficacia al programa. Generalmente, la mayoría de los sistemas de búsqueda de concordancias, de detección de errores, o de creación de ejercicios a partir de corpus, no efectúan buenas búsquedas en corpus de aprendientes, limitando el número de resultados y empobreciendo los análisis que puedan realizarse.

En este ámbito, y gracias a la bibliografía consultada, conocemos, por un lado, algunas herramientas para la mejora de la producción escrita

como *Wordpilot*[86] (Pravec, 2002), creada por Milton basándose en la experiencia con aprendientes chinos de inglés como lengua extranjera, que propone una serie de tareas y herramientas para solventar los errores como ejercicios, gramáticas interactivas o listas de palabras o expresiones poco utilizadas por los aprendientes. Y por otro lado, sistemas como el *ESL Tutor*, para ayudar a los aprendientes coreanos a evaluar sus errores más frecuentes y persistentes, y a mejorarlos o corregirlos a través de la retroalimentación proporcionada por el sistema.

En este punto podríamos incluir, además, las herramientas Exxelant (*EXample eXtractor Engine LANguage Teaching*), o MIRTO, descritas en el capítulo anterior, pero en todos los casos se trata de programas o herramientas muy ligadas a los corpus para las que fueron creadas, no siendo posible su uso generalizado con otros corpus de aprendientes, lo que, sin duda, limita enormemente su potencial.

Se necesitan aún muchos corpus de aprendientes de gran tamaño para entrenar distintos sistemas, y una colaboración estrecha entre ingenieros informáticos, lingüistas y especialistas en adquisición de segundas lenguas capaces de encontrar reglas y patrones que puedan formar parte de programas adecuados para esta variedad de lengua.

En la actualidad, se trabaja para conseguir nuevas aplicaciones específicas, muchas de ellas entorno a sistemas de creación de ejercicios basados en los errores frecuentes del aprendiente, es decir, sistemas que sean capaces de reconocer los errores, evaluarlos y proveer una cierta retroalimentación a través de ejercicios específicos para que el aprendiente pueda practicar y solventar el problema en el futuro.

5.3 Uso directo de corpus de aprendientes

El uso directo de corpus de aprendientes en enseñanza como material de aprendizaje, como venimos diciendo, no está muy extendido debido al modelo de lengua que presenta. Al menos, para su uso general con aprendientes de una L2 en contexto guiado o académico.

[86] La única información disponible, pero no el programa de descarga, se encuentra en: http://ihome.ust.hk/~lcjohn/

Existen muchas discrepancias en su uso debido al tipo de muestras que se podrían emplear para que los aprendientes puedan percibir los patrones lingüísticos y los usos frecuentes, e inducir así correctamente sus hipótesis de lengua. Se teme que no sean muestras adecuadas de lengua al contener errores y otras inconsistencias discursivas. Sin embargo, hay estudios que muestran que los propios hablantes de L2 son capaces de distinguir los errores de otros aprendientes y de intentar corregirlos.

En muchas ocasiones, se trata de encontrar al destinatario más adecuado para la tarea, un área más fácil de trabajar y los datos o corpus más apropiados para su nivel. Por ejemplo, estudios previos (Boulton, 2008, 2009; Aston et. Al, 2004) señalan que los ámbitos donde es más fácil realizar este análisis son los aspectos relacionados con el léxico (por ejemplo, uso excesivo o minoritario de ciertas palabras o expresiones, polisemia, sinonimia, *falsos amigos*, formación de palabras…) y en colocaciones simples como la complementación verbal o adjetival.

No obstante, autoras como Granger (2004) sugieren que no se debe dejar de lado el uso directo de corpus de aprendientes, ya que para aquellos que son de nivel avanzado puede tener efectos beneficiosos en la detección de errores persistentes o en la comparación de la lengua nativa y la lengua de un hablante de L2, ahondando en una mayor capacidad de autonomía de aprendizaje, de reflexión sobre la lengua y de aprendizaje por descubrimiento.

La misma opinión es compartida por Osborne (2004) que habla de proponer este tipo de análisis para cualquier tipo de aprendientes de L2, siempre y cuando sean preparados previamente por el docente. Así, Osborne considera que combinar un análisis ascendente con datos procedentes de aprendientes y un análisis descendente con muestras de corpus de hablantes nativos es quizá el uso directo de corpus más provechoso, pues permite construir muchos y variados ejercicios que pueden ayudar a los aprendientes a detectar y discriminar las discrepancias entre su uso y el de los hablantes nativos, desarrollando lo que ya conocemos por conciencia lingüística. Este trabajo con ambos tipos de corpus nos ayudará también a practicar la observación, la detección de patrones y de fenómenos lingüísticos significativos que sobresalen entre las concordancias, con el fin de deducir hipótesis y validarlas para obtener ciertas conclusiones propias sobre el uso de la lengua. No debemos olvidar tampoco que esta distinción entre el uso de

la lengua por hablantes nativos y el de hablantes de L2 puede ayudarles a detectar sus propios errores y a corregir los usos inapropiados de ciertas estructuras.

Sin embargo, y siguiendo con las pautas propuestas por Osborne (*op.cit*), muchos estudios reflejan la necesidad de llevar a cabo una metodología más estricta en el uso directo de los corpus de aprendientes para evitar que los alumnos recuerden antes la forma incorrecta e incurran en nuevos errores. Por tanto, autoras como Susan Huston (2002) y Nesselhauf (Aston et al, 2004), presentan un posible esquema de actuación que sirva para poner en práctica esta metodología, y que podría resumirse en las siguientes fases:

a) Identificación del error y comparación con corpus de nativos
b) Búsqueda de concordancias comparables entre ambos corpus que son presentadas al aprendiente para que distinga las diferencias entre el uso de un hablante de L1 y uno de L2, generalmente con ayuda de preguntas específicas o ejercicios concretos que le faciliten la observación
c) Realización y puesta en práctica de ejercicios o actividades que ayuden al aprendiente a desarrollar y fijar el uso correcto del aspecto analizado, para que no represente un error en el futuro

La inclusión de corpus de aprendientes en tareas del enfoque *data-driven learning* supone, ante todo, una ventaja importante con respecto al uso de otro tipo de corpus: el aumento del número de usuarios que pueden trabajar en esta metodología. Normalmente, las actividades propuestas (si no están, evidentemente, preeditadas por el docente) debido a la complejidad de los corpus utilizados (normalmente grandes corpus de referencia, en los que no se puede hacer una preselección de léxico y estructuras *a priori*), no son aptas para todo tipo de usuarios, siendo sólo aquellos de más nivel y conocimiento de la L2 los que son capaces de desenvolverse en las tareas.

Esta tendencia cambia radicalmente con la inserción de corpus de aprendientes, puesto que la complejidad de los textos disminuye (lo que no significa que no se enriquezca con nuevos fenómenos o usos distintos). Eso nos permite trabajar ya con alumnos que posean un nivel intermedio. En niveles más iniciales, es complicado trabajar por la forma

y el contenido de las muestras, pero, sin embargo, con una preselección y una cierta edición, también podríamos lograr realizar varias actividades más sencillas. Como venimos señalando en el presente estudio, el uso de corpus en enseñanza se puede generalizar si encontramos el corpus más adecuado para el aprendiente, o si se hace un trabajo de selección y edición previo por parte del docente.

Es cierto, además, que una gran parte de los aprendientes encuentra motivador el hecho de buscar los errores y explicar sus posibles causas. Como explica Nesselhauf (Aston et al., 2004), el hecho radica, en parte, en la dimensión afectiva. Los aprendientes ven el error en el aula como un elemento incorrecto que es sistemáticamente corregido como una mala práctica. Trabajar con errores, tener la posibilidad de identificarlos, explicarlos y corregirlos por sí mismos supone poder conseguir algo positivo asociado al error.

Además, hay diversas teorías que identifican una menor ansiedad en el aprendiente cuando analiza la lengua de otro hablante no nativo (y también en interacción oral con él), aunque este tenga una competencia superior a la suya. Habitualmente el hablante nativo es percibido por algunos aprendientes como un modelo inalcanzable, o incluso, como una especie de juez de su competencia, lo que les produce una mayor frustración. Como sabemos, una menor ansiedad, redunda en una mayor motivación ante la producción o la tarea, y por tanto, una mejor asimilación, así como una utilización posiblemente más correcta de la lengua meta por parte del aprendiente.

Por otro lado, desde hace un tiempo existe una corriente (Farr, 2008; O' Keeffe and Farr, 2003), con mayor acogida en la comunidad investigadora, que promueve la utilización de corpus de aprendientes con futuros docentes. Es decir, que preconiza su análisis y manejo en los estudios superiores de didáctica de lenguas y otras materias afines. En muchos casos, la compilación y el análisis de pequeños corpus de aprendientes locales representa la primera aproximación de muchos docentes a la investigación en enseñanza basada en corpus.

En este ámbito, se propone a los futuros docentes que realice distintos análisis de aspectos de la lengua del aprendiente para conocerla en profundidad y valorar métodos de enseñanza apropiados a sus necesidades para alcanzar una competencia efectiva en L2. Es muy frecuente que se les pida que comparen la lengua de hablantes de L2 y de hablantes de L1, o nativos (análisis contrastivos). En estos casos se suele

atender a fenómenos preponderantes como el uso distinto del léxico (con un uso excesivo o al contrario, minoritario, de algunas expresiones o colocaciones), las anomalías gramaticales más frecuentes y patrones propios del discurso de ambos.

Debemos de concluir que, tanto para el uso directo como indirecto de los corpus de aprendientes, aún queda mucho por explorar. Si bien se han realizado numerosos análisis para caracterizar la interlengua, donde los corpus de aprendientes han sido, sin duda, un elemento esencial y decisivo, todavía queda mucho por hacer en su utilización en manuales y otros materiales de referencia, y por supuesto, en el manejo automático de estos en diferentes programas de explotación de corpus, que todavía no están adaptados a la complejidad y las características de la lengua de los aprendientes.

6. Conclusiones

Considerando que la implementación de corpus de aprendientes es aún muy reciente, podemos considerar que se trata de un campo aún emergente. Las dificultades para la recolección de datos impuestas por las exigencias de un diseño muy cuidadoso y su aplicación no muy extendida hacen que no encontremos todavía un número de ejemplos significativo, sobre todo para algunos subtipos de corpus, como los de aprendientes orales o en lenguas distintas al inglés como L2, lengua pionera que ha generado la mayor parte de los corpus de aprendientes disponibles en el mercado.

Los corpus de aprendientes no hubieran tenido el desarrollo actual sin el impulso de la investigación en adquisición de segundas lenguas, que basó en la lingüística de corpus su método más extendido de análisis. Por otra parte, este avance ha servido para que los corpus de aprendientes sean el puente de comunicación entre los estudios de adquisición de segundas lenguas y el ámbito de la enseñanza, que se ha beneficiado de ambos para desarrollar herramientas mucho más útiles y provechosas. Existen así numerosos estudios a nuestra disposición que intentan matizar las características de la interlengua, sin embargo, su aplicación al margen de proyectos o estudios teóricos no es significativa.

Los corpus de aprendientes han servido para realizar análisis exhaustivos (sobre todo de errores y análisis contrastivos de la interlengua), pero no para generar aplicaciones pedagógicas de referencia. Los únicos ejemplos disponibles provienen de ciertas gramáticas y diccionarios generales y para aprendientes realizados por las editoriales encargadas de financiar los grandes corpus de referencia de aprendientes del inglés como L2. Sin embargo, no aparecen para lenguas como el español o el francés.

Su escasa aplicación proviene, en parte, de la imposibilidad de reutilización de los corpus ya implementados, bien porque son poco difundidos y quedan restringidos al laboratorio, centro o grupo investigador que lo diseñó, o bien, porque su forma y sus características técnicas no permiten su fácil explotación y reutilización.

Al contrario que para otros tipos de corpus o aplicaciones informáticas, no existe todavía un estándar para la realización de corpus de aprendientes. No se ha desarrollado ninguna metodología, ni existe un estándar para el formato de implementación de los datos (en un determinado lenguaje como XML, TEI o similar), lo que, sin duda, influye negativamente no sólo en la posibilidad de reutilización en proyectos diferentes al de su concepción, sino ya en la propia consulta del corpus, puesto que muchas de las herramientas que tenemos a nuestro alcance no son capaces de analizar correctamente los datos.

Por consiguiente, además de crear foros de discusión y de difusión para este tipo de corpus, sería muy beneficioso el poder contar con unos estándares técnicos para homogeneizar los corpus existentes y mejorar las herramientas de explotación de los mismos (Cf. Granger, 2009: 28). Por otra parte, sería también muy interesante que los docentes pudieran tener nociones de Lingüística de Corpus, de forma que pudieran crear sus propios corpus, con sus alumnos, lo cual supondría un recurso muy valioso para sus clases. Para ello, habría que cambiar en parte el perfil del docente, que podría mejorarse añadiendo la faceta de investigador, lo que redundaría sin duda en una mejora de la enseñanza.

Como hemos venido observando, no existen en la actualidad demasiados corpus de aprendientes del francés, al menos, en comparación con la gran cantidad de los existentes para el inglés. Si nos remitimos al ámbito de nuestro país, no hemos encontrado por el momento ejemplos de corpus de aprendientes de FLE cuya lengua materna sea el español. La única referencia a corpus de aprendientes en

España, la tenemos para corpus de aprendientes del inglés y del español como lengua extranjera, ambos en vías de desarrollo (CEDEL2 y C-ORALROM-ELE). Los únicos corpus que conocemos donde se incluye a hablantes de lengua materna española son los corpus multilingües como ESF Corpus[87], PAROLE[88], o FRIDA, pero en ningún caso están centrados exclusivamente en los aprendientes españoles, ni tienen entre sus objetivos una descripción de la interacción oral del aprendiente.

Por consiguiente, dado nuestro interés en contar con un análisis empírico de los errores más frecuentes de los hablantes españoles de FLE en interacciones orales espontáneas y la ausencia de materiales disponibles, optamos por crear e implementar nuestro propio corpus, CORAF (Corpus ORal de Aprendientes del Francés), que será presentado más adelante en este estudio.

[87] http://www.clarin.eu/mpi-esf-corpus
[88] Toda la información en: http://corpusdelaparole.in2p3.fr/

6. EL ANÁLISIS DE ERRORES: HISTORIA Y METODOLOGÍA

1. Introducción

Los corpus de aprendientes contienen, como sabemos, un banco de muestras de interlenguas, es decir una colección de enunciados producidos por aprendientes de una L2. Para conocer las características de dicha interlengua y saber cómo se construye, es necesario proceder a un análisis exhaustivo de los datos del corpus.

A lo largo de la historia de la Lingüística Aplicada, han surgido distintas metodologías de análisis de la interlengua, todas ellas coherentes con los enfoques y teorías predominantes en cada momento en el campo de la adquisición de segundas lenguas. Entre las técnicas tradicionales, dos destacan por haber sido las más utilizadas: el análisis contrastivo entre aprendientes y hablantes nativos, y el análisis de errores. Este tipo de procedimientos sirvió fundamentalmente para indagar el orden y el modo en que se van construyendo la gramática y las estructuras lingüísticas transitorias que componen la interlengua.

Con el paso de los años, han aparecido nuevas técnicas, que ya no se limitan a prestar atención a los errores o a las dificultades del aprendiente observables en la interlengua, sino que apuestan por identificar y medir de forma objetiva la dimensión pragmática, -relación contexto/contenido-, y la competencia lingüística, o gramaticalidad, aceptabilidad y fluidez de la producción lingüística del aprendiente. Así surgen respectivamente las modalidades de Análisis del Discurso y de Análisis de la Competencia o Actuación (*performance*).

Sumándose a los análisis clásicos, los enfoques más recientes nos proporcionan nuevas informaciones sobre la interlengua, que permiten completar su descripción y lograr una cobertura más amplia de la misma. Lo hacen apostando por medir la longitud, la complejidad y la densidad léxica del enunciado, o por desarrollar listas de frecuencias y palabras clave capaces de describir el nivel de complejidad y el tipo de uso que hace el aprendiente de la interlengua.

Por otra parte, la mayoría de los formatos de análisis, además de permitir la caracterización de la interlengua y el desarrollo de otras investigaciones asociadas, -por ejemplo, en el campo de la psicolingüística o la sociolingüística-, contienen un marcado componente pedagógico. Por consiguiente, suelen ser utilizados en el ámbito de la enseñanza y más concretamente, de la enseñanza de L2, para determinar los beneficios o los progresos de una intervención pedagógica particular[89].

A lo largo del presente capítulo, en primer lugar, aclararemos en qué consiste el análisis de datos, incidiendo especialmente en el análisis de errores de cuyo funcionamiento ofreceremos una breve panorámica, desde su origen hasta la actualidad. En segundo lugar, hablaremos de los objetivos del análisis de errores, deteniéndonos sobre los criterios más habituales que categorizan el concepto de *error*. Finalmente, describiremos cada uno de estos criterios, comentando sus taxonomías más representativas. Finalmente, dichas observaciones nos servirán de base para justificar la taxonomía que aplicaremos en la categorización de errores del corpus de aprendientes objeto de este estudio.

2. El análisis de datos: breve historia

El análisis de datos de la interlengua es una metodología frecuentemente utilizada en el ámbito de la investigación sobre los procesos de adquisición de segundas lenguas. La elección de la técnica de análisis siempre ha dependido del tipo de objetivos que se pretendían alcanzar, y del tipo de datos que se iban a manejar. En cualquier caso, puede decirse que cuatro han sido los enfoques que han dominado la historia del análisis de datos de la interlingua:

- el Análisis Contrastivo (AC);
- el Análisis de Errores (AE);
- el Análisis de la Actuación (AA);
- el Análisis del Discurso (AD).

[89] No en vano, como hemos señalado anteriormente, los corpus de aprendientes tienen de alguna manera su origen en las anotaciones que los docentes hacían en sus clases.

La tradición ha ligado cada uno de estos modelos a una determinada época, es decir que ha considerado cada uno de ellos como la superación del modelo que le había precedido en el tiempo. Sin dejar de ser cierto, a nuestro entender, los distintos modelos ya no se inscriben tanto en una dinámica de progreso lineal de la metodología, sino más bien en una perspectivas de acumulación sucesiva de enfoques que se iban complementando, permitiendo un aproximación creciente a un análisis más científico de la lengua. Nuestra afirmación se basa en que, en muchas ocasiones y cada día con más frecuencia, lo que suele hacerse es utilizar varias de estas técnicas simultáneamente.

De modo que el AE intentó superar parte de las limitaciones observadas en el AC, que vive sus horas más bajas durante los años setenta, por la irrupción de los modelos lingüísticos y psicológicos posteriores al estructuralismo y al conductismo, a saber, el generativismo y el cognitivismo. Por su parte, el AA promete completar las carencias detectadas en el AE, yendo más allá del análisis de dificultades y errores en la producción, al observar otros ítems de carácter positivo conseguidos por el aprendiente.

Lo realmente interesante de la evolución del análisis de la lengua es que se ha avanzado hacia un enfoque menos teórico, más centrado en el desarrollo de la observación empírica. Ya no se comparan hechos de lengua hipotéticos, sino que se realizan análisis rigurosos de producciones reales. Y esto se explica porque la disciplina ha pasado del interés predominantemente científico, a un interés acentuadamente pedagógico. A su vez, eso ha hecho evolucionar el concepto del error. En el caso concreto del AE, el error ya no se interpreta como una desviación de la norma lingüística que hay que corregir, sino que se considera que forma parte integrante de la interlengua, que no es correcta ni incorrecta, sino que simplemente da cuenta del estado transitorio de desarrollo del proceso de adquisición de la L2. Y ese sistema transitorio es el objeto del análisis que va a permitir mejorar la metodología de intervención y, por tanto, los resultados de aprendizaje.

Fijándose como objetivo la definición de la interlengua, uno de los primeros análisis que se realizaron fue el Análisis Contrastivo (AC). Siempre se había pensado que la lengua materna ejercía una influencia determinante sobre la adquisición de una lengua extranjera. Debido a ello, para describir el proceso de adquisición, se creía inevitable realizar una comparación entre ambas.

Pero no todos los errores observados en la producción del aprendiente se deben a una transferencia o a una interferencia de la L1. La influencia sobre los métodos y manuales de enseñanza de los enfoques y teorías psico-lingüísticas que fueron sucediéndose, resultó tener un impacto también muy elevado sobre los resultados de aprendizaje que la lengua materna (L1).

La investigación sobre la adquisición de segundas lenguas siguió su curso y, poco a poco, se pudo comprobar que los aprendientes pasaban regularmente por diferentes etapas, en las que parecían intervenir procesos universales. La universalidad se verifica, por ejemplo, cuando muchos errores específicos se repiten con cierta regularidad en un número elevado de aprendientes de un mismo nivel o estadio de aprendizaje de la L2.

De modo que se fue tomando conciencia del papel de los procesos mentales autónomos y también del papel real la L1 del aprendiente en el desarrollo de la L2. Dicha visión del aprendiente como agente activo de su propio proceso de desarrollo, fue reorientando el AC hacia el Análisis de Errores. Para ello, resultó determinante la constatación de que la observación directa de las producciones del aprendiente podía ayudar a comprender los estadios intermedios por los que pasa el aprendizaje, y también dilucidar en qué momentos la intervención de la L1 puede ser determinante, sea en sentido positivo o, por el contrario, interfiriendo negativamente en el proceso de adquisición de la lengua meta.

El AC no ha caído en desuso, pues sigue presente en muchas de las investigaciones actuales, pero se le utiliza de otra manera y con otros fines. Por ejemplo, de aparición reciente es el denominado CIA o Análisis Contrastivo de la Interlengua, que intenta superar las limitaciones del AC, partiendo de que, ciertamente, no todo el proceso de adquisición es dependiente de la lengua materna. Con CIA se realizan, por ejemplo, estudios comparativos entre interlenguas de aprendientes con distinta L1. Pero el uso de AC ha declinado en la medida en que se ha ido tomando conciencia de la pluralidad de factores que intervienen en el proceso de adquisición de una L2, además de la L1.

En el desarrollo de enfoques de análisis posteriores, destacamos en primer lugar el Análisis de la Actuación (*Performance Analysis*), que se ocupa de elementos de la producción de la interlengua, realizando en primer lugar estudios de adquisición de morfemas, siguiendo la metodología de Brown (Larsen Freeman and Long, 1991: 62) para el

análisis de la lengua materna. Todos estos estudios intentan desgranar el proceso de adquisición de morfemas gramaticales en contextos obligatorios (como ocurre con la marca del plural o la negación). De esta forma, encontramos índices sobre qué orden debe seguir el proceso de enseñanza o instrucción, y qué orden conviene que siga el aprendizaje de la lengua meta.

Más tarde, el AA se focaliza en los estudios longitudinales, a través de los cuales se preocupa por mostrar las secuencias de desarrollo de la interlengua, especificando los pasos objetivos que sigue el aprendiente para integrar un determinado conocimiento y, a ser posible, el orden en que se consuman dichas adquisiciones. Uno de los mayores descubrimientos de este tipo de análisis fue que las secuencias de desarrollo de la L1 y la L2 eran parecidas.

Posteriormente, las estrategias que utilizaba el aprendiente fueron el objeto de los estudios de desarrollo propios del AA (Larsen Freeman and Long, 1991: 67), a saber, estrategias de formación de reglas, rutinas, patrones prefabricados, imitación de expresiones estereotipadas, etcétera.

Pese a introducir nuevas técnicas, como ocurrió con el AC y el AE, el Análisis de la Actuación adolecía de ciertas limitaciones, lo que motivó que surgiera una nueva metodología que examinaba no sólo la actuación o la producción del aprendiente, sino el *input* que recibe mientras aprende. En este punto, se generaliza el Análisis del Discurso, que abarca el análisis de la conversación o de la interacción, pero sobre todo, el uso de la interlengua por el aprendiente y su adecuación a funciones y contextos concretos. Así, se realizan, entre otros, estudios y valoraciones sobre las funciones y los actos de habla, la coherencia y cohesión del discurso o las estrategias de comunicación que intervienen en una situación comunicativa. De modo que se avanza hacia una visión más pragmática de la lengua.

En la actualidad, muchos análisis se basan en este último enfoque, porque el interés por las características de la producción del aprendiente y el uso de la interlengua es mucho mayor. Sin embargo, como ya apuntamos, el resto de enfoques de análisis no ha desaparecido, ya que es habitual encontrar estudios donde coexisten varios enfoques, logrando así ofrecer una visión mucho más completa de la lengua que utiliza el aprendiente.

3. El Análisis de Errores (AE)

Analizar y explicar los errores de la interlengua ha sido una preocupación constante en el ámbito pedagógico. Todos los docentes han tratado de entender los errores de sus alumnos para poder ayudarles a mejorar su competencia en la L2.

El AE existe prácticamente desde los albores de la pedagogía. Puede considerarse que su origen se remonta a las publicaciones de los gramáticos normativos o prescriptivos del siglo XVIII (Ellis and Barkhuizen, 2005: 51). No obstante, como enfoque metodológico, sus primeras aplicaciones datan de los años 60 del siglo pasado, en que surge como alternativa al Análisis Contrastivo al que ya se le reprochaba graves limitaciones.

En la enseñanza de lenguas, el AE sirve especialmente para mejorar el aprendizaje y la instrucción, y se utiliza en dos dimensiones claves:

- Una dimensión didáctica: para contribuir a explicar, describir y corregir los errores.
- Una dimensión psicolingüística, para ayudar a comprender mejor todos los procesos mentales que se ponen en marcha a la hora de adquirir nuevas lenguas y/o lenguas extranjeras.

Entre sus aplicaciones más habituales, podemos encontrar libros que detallan los errores más frecuentes del aprendizaje, diccionarios generales de errores (como el *Longman Dictionary of Common Errors*), o bien diccionarios específicos para aprendientes.

No obstante, no podemos dejar de mencionar que el AE es uno de los enfoques de análisis más criticados. Muchos teóricos, entre los que destacan Larsen-Freeman y Long (1994), le reprochan que sólo se dedique a la observación de las producciones erróneas del aprendiente, y no saque también provecho de la observación de los aspectos positivos. Sus detractores piensan que se trata de un sistema desfasado que, al igual que el Análisis Contrastivo, conviene evitar. Un enfoque que valora al aprendiente solo en función de sus logros, limita innecesariamente el potencial del análisis. La evaluación de la competencia de un hablante de L2 ha de englobar también aspectos relativos a su fluidez, a la adecuación de su discurso, a la utilización de estrategias de comunicación

efectivas, etcétera. Dicho esto, el AE no deja de ser un interesante punto de partida para iniciar esta caracterización de la interlengua.

El error del aprendiente, pese a ser un elemento altamente informativo sobre la competencia del hablante, no siempre ha recibido la misma valoración en el ámbito de la enseñanza de las lenguas. Las diferentes teorías psicolingüísticas sobre adquisición de una lengua y los enfoques metodológicos asociados no han atribuido al error los mismos valores, ni le han otorgado el mismo papel en sus investigaciones. Podemos resumir del siguiente modo las sucesivas etapas en la valoración del error:

- Hasta finales de los años 60, el error se consideraba como algo intrínsecamente negativo, por lo que su aparición debía evitarse a toda costa y, en caso de producirse, debía hacerse lo necesario para corregirlo inmediatamente.

- La llegada del conductismo cambió en parte esta perspectiva, ya que se consideró el error como una señal de la necesidad de creación de un determinado hábito. Paralelamente, se iniciaba el estudio de las lenguas por medio del Análisis Contrastivo (AC), para el cual el error era simplemente la consecuencia de las interferencias de la L1 en el aprendizaje de la lengua meta.

- A partir de los años 70, el Análisis Contrastivo empezó a ser objeto de numerosas críticas, y algunos autores, como Corder o Richards, propusieron utilizar el error como base para el análisis de la competencia de los aprendientes. Así nació el llamado Análisis de Errores (AE o EA, en inglés), que pretendía superar las limitaciones del AC, a la vez que integrar las teorías de Chomsky sobre el desarrollo del lenguaje. Esta nueva modalidad de análisis, -que también recibiría sus correspondientes críticas en años posteriores-, preconizaba un cambio en la valoración del error, el cual pasó a convertirse en un elemento positivo, pues su estudio no solo permitía conocer mejor las etapas del desarrollo de la L2, sino que, el mejor conocimiento de los puntos conflictivos y los obstáculos del proceso de aprendizaje, ofrecía igualmente unas posibilidades renovadas de mejorar las estrategias de enseñanza.

- En los últimos años, el concepto de error ha sido el elemento esencial de explicación para enfoques como la Teoría de la interlengua y los nuevos planteamientos en torno al autoaprendizaje. De la corrección sistemática del error en busca de su erradicación progresiva, se ha pasado a una valoración del mismo como un valioso elemento definitorio de la interlengua del aprendiente. Se ha fomentado una reflexión en torno al error, y se ha estimulado la autocorrección como sistema para avanzar en la autonomía en el aprendizaje, y para favorecer el desarrollo de competencias tan esenciales como el espíritu crítico o la capacidad de aprender a aprender.

Pese a las opiniones controvertidas que ha llegado a generar, a nuestro entender, hay que valorar en su justa medida las contribuciones del AE a un mejor conocimiento de la interlengua. Esos principales méritos son:

- haber dado a conocer los múltiples orígenes del error en el aprendizaje de una L2, lo cual, recordémoslo, ha contribuido a la superación de la única explicación simplista existente hasta entonces, que era la del papel determinante de las interferencias de la L1;
- haber provocado un aumento del interés por el estudio del error del aprendiente, tanto en el seno de la lingüística, como en el desarrollo de los currícula educativos;
- haber contribuido a situar el error como uno de los indicadores del nivel de adquisición (Cf. Dulay et al., 1982: 141).

Por otro lado, las críticas al AE son sin duda también de peso. Siguiendo a Dulay et al (*op.cit*), podemos destacar las siguientes:

- La metodología descriptiva del AE resulta confusa, principalmente porque en ella se produce una identificación entre *la causa del error*, que es el producto o enunciado erróneo producido por el aprendiente, y *su explicación*, que es el proceso a través del cual se ha llegado a ese error. Según los teóricos del AE, origen y resultado de la adquisición del error

serían una misma cosa, cuando lo útil sería distinguirlos con claridad, pues el análisis del proceso de adquisición del enunciado erróneo (interacción entre procesos internos del aprendiente y el entorno) es el que puede conducirnos a conocer la causa del error y, por tanto, permitir intervenir para su erradicación.

- Las clasificaciones del AE parecen poco apropiadas para explicar los errores del aprendiente. Sus taxonomías de errores resultan excesivamente simplistas, pecan de poca homogeneidad y además, muchas de ellas destacan por su ambigüedad.

Los investigadores interesados en su potencial han tratado de superar las limitaciones del AE, introduciendo sucesivas mejoras encaminadas a la reelaborando los criterios de análisis, basados con anterioridad únicamente en la descripción lingüística. Hasta ese momento, resultaba habitual encontrarse con listados interminables de errores que los investigadores recolectaban a partir de las producciones de sus aprendientes. Se iban apuntando y acumulando datos que eran utilizados para reorientar las metodologías de enseñanza, pero no existía un interés de sistematización de la tarea, es decir que no existía la idea de crear una tipología de errores de corte científico, que dotase al enfoque de un mayor rigor objetivo.

Por otra parte, el influjo del Análisis Contrastivo seguía siendo excesivo, pues las interferencias de la lengua materna continuaban desempeñando un papel determinante en la explicación de los errores. En este punto es donde Corder (1967) introduce novedades singulares que ayudaron a cambiar en profundidad el panorama confuso en el que se movía el AE. Dicha renovación pasa directamente por la creación de una nueva metodología para el enfoque, que fue propuesta por Corder en 1974 (Corder, 1981; Ellis and Barkhuizen, 2005: 5. Las ideas de Corder no sólo permitieron una regeneración del AE, sino que introdujeron nuevos planteamientos muy importantes en torno al concepto de error, como veremos más adelante.

3.1 Computer-Aided Error Analysis (CEA)

Las nuevas tecnologías y el desarrollo de sucesivos enfoques en lingüística aplicada, cada vez más especializados, trajeron consigo cambios muy beneficiosos para el AE. Para tratar de evitar la subjetividad en los análisis, ayudar en la ardua tarea de identificación a los expertos, y utilizar la tecnología emergente, el AE tradicional ha intentado sustituirse por otro de corte automático, realizado a través de herramientas y programas informáticos. Nace así el denominado Análisis Computacional de Errores o Análisis de Errores asistido por ordenador (del inglés *Computer-based Error Analysis* o CEA).

Dagneaux y otros (1998: 164) señalan que el CEA se propone corregir las siguientes limitaciones del AE:

- Generalmente, el sistema se basa en datos heterogéneos de aprendientes;
- La categorización suele ser bastante confusa;
- No recoge todos los fenómenos de la producción del aprendiente que cabía esperar;
- Se limita al análisis de lo que el aprendiente no sabe hacer;
- Proporciona una imagen bastante estática del aprendizaje de la lengua meta.

Por su parte, el CEA se caracteriza por contar con la ventaja inestimable de trabajar sobre datos almacenados digitalmente (procedentes en su gran mayoría de corpus escritos y orales, que ya empiezan a generalizarse en los años 90), y por utilizar herramientas y aplicaciones informáticas para la detección, edición y reutilización posterior del error. El cambio es importante desde el punto de vista cuantitativo y cualitativo, pues los análisis se apoyan ahora en un volumen de datos más elevado y en una mayor rigurosidad en la homogeneización y categorización de los errores.

Sin embargo, este tipo de herramientas de CEA no están todavía muy extendidas, y en la mayoría de los casos, permanecen en el marco específico de ciertos proyectos de investigación. La detección automática de errores es, en este momento, todavía una asignatura pendiente, debido, principalmente, a la complejidad de la lengua que se ha de procesar. La mayoría de las herramientas existentes sólo obtienen buenos

resultados con muestras de lengua procedentes de nativos y con categorías lingüísticas muy concretas, como las cerradas, donde existe una menor posibilidad de variedad de respuestas, y pueden controlarse fácilmente todos los usos. Debido a ello, la mayoría de editores disponibles no llegan a ser difundidos, siendo desconocidos fuera del ámbito de los proyectos de investigación que los vieron nacer.

Sin embargo, el CEA ofrece unas ventajas notables como que permite trabajar con aplicaciones que utilizan etiquetadores automáticos (o semi-automáticos) de categorías de errores, que van aplicando reglas de forma jerárquica, obteniendo, por ejemplo, buenos resultados en la detección de errores en la conjugación verbal de producciones escritas (llevados a cabo, por ejemplo, por Granger en su proyecto de corpus de inglés L2, *ICLE*).

Los mejores resultados que se han alcanzado con el CEA están en la descripción de las categorías lingüísticas, ya que es la tarea más sencilla para este tipo de editores. El resto de funcionalidades (como el análisis etiológico y la comparación de la interlengua con la lengua meta) todavía no han sido muy desarrolladas, debido a que en su proceder está presente una cierta subjetividad que el programa no es capaz de desentrañar.

Otro de los aspectos positivos del CEA es la incorporación de análisis estadísticos y cuantitativos mucho más rápidos y fiables, que enriquecen el análisis de errores y le dotan de una mayor objetividad, produciendo por tanto resultados también más efectivos.

En cuanto a la utilización posterior del error, existen sistemas que proporcionan una respuesta o retroalimentación al aprendiente en forma de sugerencias de reglas gramaticales que han de revisar, e incluso, en el caso de aplicaciones más desarrolladas, aparecen propuestas de ejercicios para practicar las reglas que el aprendiente no domina, contribuyendo el propio sistema a solucionar el error, y evitando así que se *fosilice*.

El mayor problema es que estas aplicaciones no son reutilizadas fuera del proyecto en que fueron concebidas, pues no son distribuidas ni puestas a disposición de la comunidad científica.

Así las cosas, y aunque se sigan desarrollando versiones informatizadas del análisis de errores, todavía queda por hacer al investigador mucho trabajo manual. En la mayoría de los casos, esas tareas consisten en etiquetar o anotar manualmente los textos o

transcripciones de los corpus (aunque con la ayuda de herramientas informáticas o editores como *TagEditor* o *UCLEE*, *Université Catholique de Louvain Error Editor,* que, por otra parte, no son fáciles de conseguir), como muestran numerosos corpus de aprendientes, como el citado NICT JLE, *Cambridge Learner Corpus*, FALKO, JFELL, *Freetext*, ICLE, o FRIDA. En el caso de este último, como ya hemos comentado anteriormente, se cuenta además de con el editor de errores, con un sistema denominado *Exxelant*, capaz de proporcionar distintos ejercicios en función del error detectado.

Por otra parte, la mayoría de los AE y CEA, como es constatable, se han desarrollado en el campo de la adquisición de segundas lenguas, destacando el inglés, como es habitual en el ámbito de la LC. En la mayoría de aplicaciones de AE o CEA para las distintas lenguas, y entre ellas, del francés, encontramos análisis focalizados, sobre todo, en la expresión escrita. La habilidad lingüística se analiza a través de la producción escrita, en parte, porque es más fácilmente cuantificable que el resto. Esta destreza ha sido evaluada desde todos los puntos de vista, a través de corpus de producciones escritas de aprendientes que abarcan distintos géneros como los tests, cuestionarios, composiciones dirigidas, cartas, etcétera, con el fin de establecer una evaluación de varios registros de lengua: formal e informal, espontánea, académica…

Sin embargo, es mucho más difícil encontrar estudios basados en corpus orales, y aún más, que describan el francés como L2. Encontramos sólo algunos precedentes en corpus orales de aprendientes para el inglés y el español. En el ámbito de FLE existe, casi exclusivamente, el corpus FLLOC de la Universidad de Southampton, dirigido por Florence Myles (aunque no se trata de un análisis de errores asistido por ordenador) y el proyecto europeo *Freetext*[90].

En el campo de la lengua escrita, también cabe destacar las investigaciones de Sylviane Granger en la Universidad de Lovaina, quien realiza corpus escritos anotados con errores tanto para el inglés (ICLE) como para el francés (FRIDA), utilizando para ello un análisis informático riguroso de errores, y creando programas de edición y análisis bastante efectivos.

Finalmente, hemos de lamentar que pese a los intentos notables de establecer un AE mucho más objetivo y fiable, algunas de sus

[90] Disponible en <http://www.latl.unige.ch/freetext/index.html>.

debilidades no han podido ser solucionadas. Por un lado, persiste una excesiva heterogeneidad en la elaboración y aplicación de las taxonomías de errores, pues hay prácticamente tantas taxonomías como proyectos o corpus desarrollados. Por otro lado, la confusión respecto a la identificación de los errores y su aislamiento de las causas que los provocan, sigue en pie, pues se continúa encontrando serias dificultades para aplicar las categorías o para determinar sin vacilar si se trata de un error de competencia o de producción.

En lo sucesivo, sería recomendable apostar no sólo por la universalización de las herramientas de edición, detección y análisis de errores, sino también por una estandarización de las taxonomías de errores, lo que facilitaría la comparación entre lenguas y aprendientes, y contribuiría, sin duda, a una mejor valoración del AE.

4. Consideraciones en torno al concepto de error

En la investigación sobre la adquisición de lenguas y, sobre todo, en el análisis de la lengua del aprendiente, el concepto de error ha sido uno de los primeros en analizarse. Con el término *error* estamos haciendo referencia a los rasgos o aspectos de la producción del aprendiente (tanto oral como escrita) que se desvían del uso habitual de la lengua de estudio por parte de los hablantes.

Generalmente, se asocia de forma abusiva y reduccionista el término de error con la idea de incorrección. Un error suele ser visto como un fracaso, un *output* inadecuado de la lengua meta. Se podrían hacer algunas concesiones a esta perspectiva, sin embargo, lo relevante tanto desde el punto de vista de la investigación como de las prácticas educativas, es que el error es fundamentalmente una marca propia de la lengua del aprendiente, un rasgo distintivo, cuyo conocimiento resulta muy productivo tanto a efectos de estrategia de enseñanza, como de aprendizaje, pues nos revela una información muy valiosa sobre las dificultades del proceso de adquisición de la L2 y sobre el camino que recorre el aprendiente hasta perfeccionar su competencia. Un error es, sobre todo, la evidencia de que hay un proceso de adquisición en marcha, y ese es precisamente el objetivo central de la instrucción.

No todas las producciones distintas o no conformes a las reglas de la gramática de la lengua meta que generan los aprendientes deben ser catalogadas como un *error*. Es muy interesante tener en cuenta en este punto la distinción tradicional entre *error* y *falta* o *errata* establecida por Corder (1967). Para este autor, es muy importante efectuar una distinción entre un hecho sistemático y otro esporádico.

A la hora de producir un discurso en cualquier lengua, tanto si es L1 como es L2, nos encontramos con enunciados no conformes a la norma en uso, que pueden sobrevenir por razones de lapsus, falta de atención, cansancio, etcétera. Si esas variantes singulares son efectivamente producto de unas condiciones psicológicas o emocionales particulares, y no dan cuenta de un déficit en el conocimiento de la lengua por parte del hablante, estamos ante lo que Corder considera una *falta*, que en inglés se denomina *errores de producción* o *ejecución* (*errors of perfomance*), y nosotros calificamos de *errata*. Los citados fallos pueden incluso ser (auto)corregidos con una mayor o menor espontaneidad y rapidez por parte del aprendiente, lo que de hecho revelaría que es consciente de los mismos. La otra modalidad es denominada por Corder *errores de competencia* (*errors of competence*), y son aquellos fallos que sí se producen de manera sistemática, denotando, por tanto un déficit evidente en la competencia del aprendiente.

El principal problema es que, en la práctica, no siempre resulta fácil distinguir si el fallo del aprendiente es catalogable como error o como errata, sobre todo si lo analizamos a través de producciones espontáneas aisladas. Evidentemente, las erratas no son significativas para el proceso de aprendizaje de una lengua. Es solo el error el que aporta una gran cantidad de información que puede ser utilizada con fines pedagógicos.

Según Corder (1967), el error conlleva una triple utilidad:

1. Para el profesor, porque el análisis sistemático de errores le proporciona información sobre el camino recorrido por el aprendiente y el que le queda por recorrer en su proceso formativo.

2. Para el investigador, porque le proporciona evidencias sobre el proceso de adquisición de una lengua y, más concretamente, sobre las estrategias internas de los aprendientes.

3. Para el propio aprendiente, porque puede apoyarse en el error para verificar sus hipótesis sobre el funcionamiento de la lengua que está adquiriendo.

Pero el concepto de error sigue siendo objeto de controversias. Para algunos expertos como Jenkins (Cf. Aijmer, 2009: 25), los términos *interlengua* o *lengua del aprendiente*, incluso el de *error*, no resultan acertados, porque la L2 no debe considerarse una lengua distinta a la lengua meta, sino una realización diferente de esta. Es decir, una variedad más de la lengua, o un *registro* con sus rasgos distintivos correspondientes, tan apropiados y legítimos como los de cualquier otro registro de la lengua meta (variedades regionales, lenguas de especialidad, lenguaje común, sostenido, infantil, vulgar…). Obviamente, desde este punto de vista, el análisis contrastivo con hablantes nativos sería inadecuado. Los expertos recurren entonces al análisis de patrones léxicos y sintácticos frecuentes, o de cambios en el uso de los mismos, con el propósito de ampliar los indicadores de aceptabilidad basándose en una muestra más abierta de los usos de la lengua. Los enunciados singulares ya no serían un reflejo de desviaciones de la lengua normativa, sino más bien la manifestación de un uso particular.

Analizar un error, como puede verse, no es tarea sencilla, pues es frecuente caer en interpretaciones poco acertadas del mismo. Como ya hemos comentado, no resulta fácil en muchas ocasiones distinguir si estamos ante un error de competencia o ante un fallo aislado que no se repite en otros contextos. Además, muchos fallos pueden resultar difíciles de catalogar por englobar más de un error en la misma forma.

Por otra parte, en el mundo de la enseñanza de L2, se está generalizando el principio según el cual no se trata de que el aprendiente utilice correctamente una lengua en el sentido normativo, sino de lograr que su competencia de comunicación sea eficaz en diversos los contextos. En este punto, al analista de errores le surge un grave dilema, puesto que actualmente, para determinar si una forma debe catalogarse o no como error, nos guiamos por una norma, generalmente muy rígida, de la lengua escrita, y propia de un contexto formal o, digamos que asociable a una competencia avanzada en la L2. Es inevitable pensar que operando como lo hacemos, podemos estar incurriendo en la falacia comparativa, porque probablemente, en una situación de comunicación real, esos mismos errores no ocasionarían problemas en la comunicación

con hablantes nativos, pues con frecuencia el contexto hace posible que el error pase desapercibido.

De hecho, en la literatura contemporánea sobre adquisición de lenguas, se considera que la corrección académica con la que un hablante de L2 se expresa no es el único indicativo, ni siquiera el más representativo de una competencia avanzada en la lengua meta. Existen otros indicadores determinantes como la fluidez, el uso de expresiones fijas o colocaciones adecuadas al contexto de la situación comunicativa, el dominio de elementos discursivos y pragmáticos, o de diferentes registros de una misma lengua. Es posible que el futuro del análisis de errores y del análisis de corpus de aprendientes nos depare el desarrollo de una tecnología de nueva generación que permita contemplar una lengua de comunicación en todas sus dimensiones. De momento, los investigadores están en ello.

5. Análisis de Errores: descripción y explicación de los errores

Como queda explicado anteriormente, un valor esencial del AE es que sea sistemático, con el fin de que pueda ser considerado válido y representativo. Por tanto, es primordial haber establecido con antelación una taxonomía de los errores que van a ser analizados para cada una de las áreas de estudio. Esta tarea se conoce como AE deductivo, lo que simplemente significa que aquello que se va a analizar ha quedado establecido con anterioridad. En nuestro estudio, hemos apostado precisamente por un AE deductivo. El caso contrario, sería el de un análisis de errores inductivo, es decir, aquel en el que primero se identificarían los errores, y después se procedería a su clasificación.

Un problema que presenta el AE es que todavía no se ha conseguido llegar a una taxonomía compartida por todos los proyectos de este tipo. Se le reprocha efectivamente el no contar con un paradigma de estudio claro. La citada confusión se explica por el hecho de que cada proyecto conlleva la necesidad de contemplar determinados aspectos por encima de otros, en función de su naturaleza y de su finalidad específica. Debido a ello, y siguiendo la metodología más habitual, intentaremos

describir aquí brevemente algunas de las taxonomías más utilizadas, para después justificar la metodología utilizada en nuestro propio estudio.

Pese a la dispersión, la mayoría de los enfoques coinciden en utilizar tres criterios principales para el desarrollo de sus análisis:

- El criterio lingüístico
- El criterio descriptivo
- El criterio etiológico

Del *criterio lingüístico* podemos decir que es meramente informativo, ya que alude a la categoría que se encuentra afectada por el error. Se señala la parte de la oración a la que corresponde, pero no se va más allá. Por lo tanto, es efectivo a niveles estadísticos, pero no proporciona aspectos interesantes para la definición del proceso o el grado de adquisición en el que se encuentra el aprendiente.

El *criterio descriptivo* trata de profundizar más en el análisis de la forma errónea, describiendo el mecanismo de cambio que ha tenido lugar en la interlengua, y el proceso que el aprendiente ha seguido hasta llegar a producir esa forma concreta. En este caso, el AE proporciona más información, no sobre el grado de adquisición del aprendiente, sino proporcionando datos orientados a la corrección de errores. Además, el conocimiento del mecanismo de cambio producido en la interlengua sirve de ayuda para el siguiente análisis, el etiológico, ofreciendo también orientaciones sobre la causa de la aparición de dicho error.

El *criterio etiológico* es quizá el más significativo, ya que aporta información muy valiosa sobre el grado del proceso de adquisición en el que se sitúa el aprendiente. Con este análisis, podemos validar nuestras hipótesis sobre el grado de adquisición del aprendiente y establecer por consiguiente, mejoras y aplicaciones para que el aprendizaje sea mucho más rápido y efectivo.

Como ya hemos observado antes, casi todos los enfoques de AE son distintos en función de aquello que el/los investigador/es quiera/n remarcar. La mayoría de los proyectos realizan un AE basado en estos tres criterios iniciales, pero el desarrollo de nuevas disciplinas y otras corrientes de enseñanza, ha propiciado la aparición de nuevos criterios

de análisis de las producciones orales o escritas del aprendiente. Actualmente, existen análisis de errores que incluyen además criterios comunicativos, que informan sobre la efectividad de la interacción, o criterios pragmáticos y culturales, que ahondan en un conocimiento de la lengua que no se basa sólo en niveles lingüísticos, sino que percibe a ésta como un conjunto de saberes necesarios para la comunicación más allá de la gramática.

6. Metodología del Análisis de Errores: Taxonomía para la categorización de errores

La metodología propuesta por Corder (Ellis and Barkhuizen, 2005), ya comentada en el presente capítulo, presenta las siguientes fases o procesos:

1. Recopilación de muestras de lengua de los aprendientes
2. Identificación de los errores
3. Descripción
4. Explicación
5. Evaluación de los errores
6. Proposición de actuaciones para la resolución de los errores (si el AE se enmarca dentro de una perspectiva didáctica o pedagógica

Tras realizar la recopilación del corpus, y una vez elegidos los criterios desde los que vamos a analizar los errores, lo procedente es rastrear las taxonomías disponibles que más se adaptarían a las necesidades de nuestro estudio. Con ello, lo que hacemos es contribuir a la estandarización de los AE que ya existen

En nuestro caso, buscamos un tipo concreto de clasificación de errores, que nos permita dotar nuestro AE de cierta sistematicidad y objetividad. Una taxonomía adaptada nos ayudará a realizar análisis estadísticos y cuantitativos, y comparar los resultados con los de otros estudios sobre aprendientes de distintas lenguas que utilizaron la misma taxonomía de errores.

A continuación, realizamos hacemos un repaso crítico de las más difundidas que utilizan los tres criterios que queremos aplicar a los datos de nuestro corpus: criterio lingüístico, descriptivo y etiológico.

6.2 Criterio lingüístico

El primer análisis que podemos hacer del *output* del aprendiente consiste en determinar la categoría gramatical que se ve afectada por el error. Así, distinguiremos entre verbo, nombre, artículo, pronombre, preposición, etcétera, entre todas las partes que componen la oración. Generalmente se trata de identificar también a qué plano del discurso afectan estos errores, destacando si se trata de un error a un nivel léxico, semántico, morfosintáctico, discursivo…

Las taxonomías de errores que encontramos dentro de este tipo de análisis son mucho más homogéneas que en el resto, principalmente porque aluden a una estructura de sobra jerarquizada como son las partes de la oración. Además, muchos de los análisis de lengua de aprendientes estudiados, como los de los corpus ICLE, *FreeText Project*, *Cambridge Learner Corpus*, NICT JLE, FRIDA o LINDSAI, se basan en análisis automáticos de errores, realizados a través de complejos programas informáticos que implican etiquetadores automáticos, gramáticas, etcétera, por lo que los campos están conscientemente delimitados para evitar muchas ambigüedades (aunque, por supuesto, este tipo de programas nunca acaba por librarse del todo de errores provocados por la ambigüedad intrínseca a algunas estructuras gramaticales).

Para nuestro estudio, tomaremos como base la anotación automática de errores propuesta por Sylviane Granger (2003a) en su corpus escrito FRIDA (*French Interlanguage Database*). Nuestra intención es que los datos que obtengamos se puedan comparar y reutilizar. Además, el hecho de contemplar las mismas etiquetas podría favorecer el uso de algunas de las herramientas que utilizaron en FRIDA.

Para adaptar el modelo a nuestras necesidades, realizaremos pequeñas adaptaciones en la taxonomía. En primer lugar, al no trabajar con producciones escritas, nos sobrará todo aquello que se refiere a la ortografía y la puntuación. Sin embargo, nos interesa mostrar los errores propios de una interacción oral, principalmente fonéticos, por lo que

añadiremos una nueva etiqueta a la taxonomía para señalar un error fonético (marcado con <PHO>). Somos conscientes también de que es posible que encontremos errores que no sepamos bien a qué categoría pertenecen, por tener límites difusos o porque al tratarse de una producción oral, puede haber errores que se deban a una deficiente competencia en expresión oral[91]. Para todos estos casos hemos introducido la etiqueta <AMB>, que agrupa los errores que pueden verse afectados por más de una categoría o sobre los que tenemos serias dudas con respecto a su clasificación.

La taxonomía de categorías de errores para el criterio lingüístico puede resumirse según la siguiente tabla:

CAMPOS DE ERROR		CATEGORÍAS DE ERROR	
<M>	Morfología	<MDP>	Derivación-Prefijación
		<MDS>	Derivación-Sufijación
		<MFL>	Flexión
		<MFC>	Confusión de flexión
		<MCO>	Palabras compuestas
<G>	Gramática	<CLA>	Clase
		<AUX>	Auxiliar
		<GEN>	Género
		<MOD>	Modo
		<NBR>	Número
		<PER>	Persona
		<TPS>	Tiempo verbal
		<VOI>	Voz

[91] Es muy frecuente que tengamos dudas, por ejemplo, a la hora de señalar errores fonéticos en los que se vean implicados los artículos definidos *le* y *les*, o adjetivos de género femenino. En estos casos somos conscientes de que existe un error fonético, pues la pronunciación no es correcta, pero no podemos asegurar que se superponga un error gramatical, relacionado con la ausencia de concordancia entre género y/o número.

<L>	Léxico	<SIG> <CPA> <CPD> <CPV> <CPN>	Significado (incluye creación léxica) Complementación adjetival Complementación adverbial Complementación verbal Complementación nominal
<X>	Sintaxis	<ORD> <MAN> <RED>	Orden de palabras Palabra necesaria omitida Palabra innecesaria / redundante
<PHO>	Fonética	<PHO>	
<AMB>	Ambiguos	<AMB>	

Tabla 3: Clasificación de errores lingüísticos (Adaptado de Granger, 2003a: 468)

Granger (2003a) establece asimismo las siguientes etiquetas correspondientes a categorías gramaticales, que resultan igualmente útiles para nuestro estudio:

CATEGORÍA GRAMATICAL			**CATEGORÍA GRAMATICAL**		
ADJETIVO	SIMPLE	<ADJ>	PRONOMBRE	DEMOSTRATIVO	<POD>
	COMPARATIVO	<AJC>		POSESIVO	<POP>
	SUPERLATIVO	<AJX>		PERSONAL	<POO>
	COMPLEJO	<AJL>		INDEFINIDO	<POI>
ADVERBIO	SIMPLE	<ADV>		EXCLAMATIVO-INTERROGATIVO	<POX>
	COMPLEJO	<AVL>		NUMERAL	<PON>
ARTÍCULO	DEFINIDO	<ADE>		ADVERBIAL	<POA>

	INDEFINIDO	<AIN>		RELATIVO	<POR>
	PARTITIVO	<APA>		IMPERSONAL	<POS>
	CONTRACTO	<ACO>			
CONJUNCIÓN	COORDINACIÓN	<COC>	NOMBRE	SIMPLE	<NOM>
				COMPUESTO	<NOC>
	SUB. SIMPLE	<COS>		COMPLEJO	<NOL>
	SUB. COMPUESTA	<COL>		PROPIO	<NOP>
PREPOSICIÓN	SIMPLE	<PES>	VERBO	SIMPLE FINITO	<VSC>
	COMPUESTA	<PEL>		PARTICIPIO SIMPLE	<VSP>
DETERMINANTE	DEMOSTRATIVO	<DED>		GERUNDIO SIMPLE	<VSG>
	POSESIVO	<DEP>		INFINITIVO SIMPLE	<VSI>
	INDEFINIDO	<DEI>		FINITO COMPUESTO	<VCC>
	EXCLAMATIVO-INTERROGATIVO	<DEX>		PARTICIPIO COMPUESTO	<VCP>
	RELATIVO	<DER>		GERUNDIO COMPUESTO	<VCG>
	NUMERAL	<DEN>		INF. COMP.	<VCI>

Tabla 4: Categorías gramaticales que pueden verse alteradas por los errores del aprendiente (adaptado de Granger, 2003a: 479).

Pese a nuestro intento de utilizar una taxonomía operativa, correctamente desarrollada, y adaptable a nuestro estudio, al realizar el análisis, nos dimos cuenta de que los errores susceptibles de ser etiquetados bajo varias categorías eran más numerosos de lo previsto. La distinción de los errores entre gramaticales, sintácticos y morfológicos, y léxicos (denominaciones ya confusas de por sí[92]), que propone Granger (2003a), en la práctica, no resulta tan evidente como parece. Por ejemplo, algunos errores morfológicos pueden considerarse también como léxicos y, evidentemente, muchos errores morfológicos afectan también a la sintaxis. No obstante, hay que admitir que la taxonomía de descripción por categorías gramaticales es bastante precisa, ya que recoge el conjunto de categorías de forma suficiente, clara y ordenada.

Hemos elegido esta taxonomía como punto de partida para nuestro estudio, precisamente por lo minucioso de sus descripciones de las partes de la oración. Sin embargo, en cuanto a los criterios generales lingüísticos del error, creemos que sería necesario introducir algunos cambios, o delimitar algo más las características de cada uno de los subtipos, ya que, como apuntamos, muchos errores podrían encajar en varios tipos.

6.2 Criterio descriptivo

En este nivel de análisis, se pretende mostrar qué mecanismo ha provocado el cambio en el enunciado oral del aprendiente (*output*). Para ello, comparamos la forma de la interlengua y la considerada correcta desde un punto de vista normativo. La descripción del error muestra el resultado de la adquisición y como en el caso anterior, es un análisis superficial que se basa, ante todo, en la forma y las características externas del *output* en la L2.

A la hora de realizar este análisis, tenemos que tener cuidado con el tipo de comparación que vamos a establecer con la forma de la lengua meta. Debemos tener siempre presente qué es lo que el aprendiente quiere decir, adaptando nuestras expectativas al nivel de dominio de la lengua del aprendiente, es decir, evitando comparar el enunciado con

[92] Cf. Tabla 3

formas demasiado complejas de la lengua meta, que él no está preparado para producir.

También tenemos que ser conscientes de qué variedad y norma de la L2 debemos comparar con el *output* del aprendiente. En nuestro caso, como ya hemos indicado, nos encontramos frente a un registro oral propio de conversaciones en tono familiar. Como se sabe, algunos rasgos de la lengua oral son agramaticales desde el punto de vista de la lengua escrita, pero no por ello deben catalogarse como errores, ya que para la lengua oral, sí resultan adecuados: en el caso del francés, afecta, por ejemplo, a la ausencia de elementos de la negación, uso de la negación entonativa, uso de pronombre '*on'* con verbo conjugado en tercera persona de singular con intención de plural, repetición de sujeto por *dislocation à gauche*[93]….

En la literatura sobre AE se han distinguido varias taxonomías, pero, a diferencia de lo que ocurre con los criterios lingüístico y etiológico, existe aquí un mayor consenso en torno a las subcategorías.

Corder (*op.cit.*) fue uno de los primeros en proponer criterios descriptivos aplicables a los errores de aprendientes. Propuso las siguientes distinciones:

- *Omission*: omisión de elementos obligatorios en la producción del aprendiente;
- *Addition*: adición de elementos que son innecesarios para la gramaticalidad de la oración o que son incorrectos;
- *Selection of an incorrect element*, es decir, la selección de un elemento que no es el que correspondería en el enunciado (en estudios posteriores se tradujo por *falsa selección*);
- *Misordering of elements*, u orden inadecuado de los elementos que componen el *output* (también denominado por otros autores como *orden incorrecto*).

Posteriormente, Dulay, Burt y Krashen (1982: 150) señalan una tipología muy similar para explicar los mecanismos que intervienen en las formas erróneas producidas por los aprendientes, denominada *Surface*

[93] Cuando se dice, por ejemplo: "Qui est-ce, Bernanos?" en lugar de "Bernanos, qui est-ce?".

Structure Taxonomy. Esta taxonomía intenta reflejar el modo en que la estructura de la forma meta cambia en las producciones erróneas. Se habla entonces de cuatro grandes cambios en la estructura:

- *Omission* (omisión)
- *Adition* (Adición), que incluye: *regularization, double-marking* y *simple aditions*.
- *Misinformation* (forma errónea), que abarca los procesos de: *regularization, archi-forms* y *alternating forms*.
- *Misordering* (colocación falsa).

Posteriormente, James (1998) retoma el marco teórico de AE iniciado por Corder y apoyándose en el trabajo de Dulay et al. (1982) en la *Surface Structure Taxonomy*, propone sus propias estrategias de cambio en la llamada *Target Modification Taxonomy*. Así, distingue entre errores de:

- Omisión
- Adición
- Elección falsa
- Colocación falsa
- Asociación Cruzada (que denomina *blends*): Supone una innovación con respecto a la categoría de Dulay y otros, y pretende tipificar aquellos errores en los que encontramos una mezcla de dos tipos gramaticales[94].

Resumiendo las taxonomías anteriores, y tomando, sobre todo, elementos de las categorías descritas por Vázquez (1991) y Santos Gargallo (1993) para ELE (Español Lengua Extranjera) coincidentes con la taxonomía de James, nuestra propuesta es explicar los cambios que se producen en las producciones propias de hablantes no nativos a través de las siguientes categorías:

[94] Quizá esta categoría sería más pertinente para el criterio etiológico, pues explica el origen del error, sin embargo James lo mantiene entre aquellas categorías propias de un análisis descriptivo.

- **Omisión:** Explica la ausencia de un morfema o de una palabra que debería aparecer para que la estructura de la oración sea funcionalmente correcta. Por ejemplo, '**je ne parle bien*', donde se olvida parte de la negación obligatoria.

- **Adición:** Supone la presencia de un morfema o de una palabra, que, no es necesario en dicha estructura, y suele constituir un elemento redundante. Lo podemos observar en: "**les français sont* être *en silence*".

- **Elección falsa o Falsa selección:** Utilización de un morfema, palabra o estructura de forma incorrecta para un determinado contexto. Lo más normal es que aparezca una unidad gramatical distinta a la que sería necesaria o es habitual en ese contexto, por ejemplo, en el caso de las categorías cerradas de palabras, como las preposiciones: '*j'ai étudié cette langue dans l'école*', en vez de '*j'ai étudié cette langue à l'école*'.

- **Forma Errónea**: Aparición de una forma errónea dentro de un morfema o en una estructura en la que no debería de utilizarse. Generalmente suelen ser errores derivados de una incorrecta aplicación de las normas gramaticales. (Por ejemplo, *j'*étude*, en vez de *j'étudie*).

- **Colocación Falsa / Orden Incorrecto**: Utilización de un morfema o conjunto de morfemas de manera que el orden sintagmático apropiado resulta alterado, y por tanto, es incorrecto. Se puede observar en: '*je n'ai pas un souvenir bon de Paris*', donde *bon* tendría que situarse antes del sustantivo.

6.3 Criterio etiológico

El criterio etiológico ofrece el análisis más amplio y profundo que podamos realizar de las producciones de los hablantes de una L2. A través de él, pueden conocerse las causas que motivan un determinado comportamiento lingüístico, y se pueden realizar hipótesis sobre el origen de los cambios y de las desviaciones de las producciones del aprendiente en relación con el uso de la lengua por parte de hablantes nativos. Este procedimiento mantiene una estrecha relación con aspectos

psicolingüísticos que intervienen en el aprendizaje de una L2, y a través de él, se intenta descifrar el proceso subyacente que interviene en el aprendiente a la hora de apropiarse la L2.

La aparición de un error es la señal de que existe una deficiencia en la competencia del aprendiente. En muchas ocasiones, estos errores proceden de una necesidad para comunicarse con elementos de la L2 aún desconocidos por él. En ese momento, el aprendiente moviliza estrategias cognitivas alternativas que le permiten superar el obstáculo de comunicación causado por una insuficiencia en su competencia. El análisis etiológico permite indagar el origen de esa falla en la competencia lingüística.

Muchos son los autores que se han interesado por este tipo de descripción, siendo el criterio más estudiado de los tres que venimos comentando. Entre los primeros autores que propusieron una clasificación de las causas de los errores está Selinker (1972), quien describe cinco posibles procesos cognitivos en el aprendiente:

- Transferencia lingüística (positiva y negativa)
- Transferencia de instrucción
- Estrategias de aprendizaje
- Estrategias de comunicación
- Hipergeneralización de reglas de la L0

Por otra parte, Richards (1971a) describe tres categorías básicas de errores, ahondando en su tipología y en los posibles mecanismos de interferencia, resultando una clasificación basada en:

- Errores Interlinguales;
- Errores Evolutivos, que son los que reflejan el proceso de aprendizaje del hablante no nativo, y que se producen, básicamente, porque intenta construir hipótesis sobre el funcionamiento de la L2 valiéndose del conocimiento que va adquiriendo, que aún resulta insuficiente.

- Errores Intralinguales, que pueden ser de tres tipos:
    - Hipergeneralización: El aprendiente generaliza una regla que conoce de forma indiscriminada y la aplica en contextos o paradigmas donde no es aplicable, simplemente porque asimila esas formas a otras parecidas. Dicha hipergeneralización puede tener dos causas:
        - Hipercorrección: Mecanismo por el cual se tiende a inferir una excepción como una regla para un determinado paradigma, es decir, la excepción se convierte en la regla a seguir en contextos similares.
        - Sobreaprendizaje (*Overlearning*) de una estructura: Errores que provienen de una instrucción deficiente o resultado indeseado de ciertas técnicas de instrucción.
    - Ignorancia de las restricciones de las reglas: Errores producidos porque el aprendiente no es capaz de aplicar convenientemente aquellas reglas que ya conoce, sobre todo, siendo consciente de las limitaciones o casos de uso específicos en los que no se puede aplicar. Es muy parecida a la generalización, puesto que el aprendiente se sirve de un conocimiento previamente adquirido y lo aplica a casos que aún no conoce.
    - Aplicación incompleta de las reglas: El aprendiente no es capaz de aplicar reglas que ya conoce y que existen en la interlengua, tendiendo a una simplificación del sistema.

Dulay y otros (1982:163) también son conscientes de la importancia del criterio etiológico para describir los errores, y proponen una categoría llamada *Comparative Taxonomy*, basada en cuatro tipos:

- Errores de interferencia o interlinguales: Aquellos que aparecen por la interferencia en el aprendizaje de la estructura propia de la L1.
- Errores evolutivos: Errores que se derivan del proceso de desarrollo de la interlengua.

- Errores ambiguos: Errores que pueden pertenecer a las dos categorías anteriores.
- Errores únicos u otros errores: Aquellos errores que no pueden ser tipificados como evolutivos o como interlinguales.

Posteriormente, James (1998), que ya había realizado la ya mencionada *Target Modification Taxonomy*, complementa su análisis de errores con cuatro amplias categorías:

- Errores interlinguales, en su mayoría, por una transferencia negativa o interferencia de la lengua materna y otras lenguas conocidas.
- Errores intralinguales, que aparecen porque el aprendiente intenta suplir sus carencias de competencia a través de distintas estrategias de comunicación y de aprendizaje. Suele ser la categoría más estudiada por su interés pedagógico y James la subdivide en siete subtipos:
    - Falsas analogías o asociaciones cruzadas
    - Hipótesis falsas
    - Aplicación incompleta de las reglas
    - Redundancia
    - Ignorancia de las restricciones de las reglas
    - Hipercorrección
    - Hipergeneralización o simplificación
- Estrategias de comunicación, generalmente válidas para explicar muchos de los errores léxicos que aparecen en el *output* del aprendiente.
- Errores inducidos: aquellos que aparecen debido a la instrucción que el aprendiente ha recibido por una explicación inadecuada de la gramática por parte del profesor, de los materiales didácticos con los que trabaja, por una asimilación errónea del aprendiente, entre otros.

Como ya queda comentado, es en el ámbito del Español como Lengua Extranjera donde hemos encontrado más categorizaciones etiológicas de errores, que además se asemejan mucho más entre ellas que en el caso de otras lenguas[95]. Hay que destacar las utilizadas por Vázquez (1991) y por Fernández López (1997). Por un lado, Vázquez (1991) alude a tres tipos básicos de errores:

- Los errores interlinguales
- Los errores intralinguales
- Los errores evolutivos, que pueden tener relacionarse con mecanismos de:
  - Analogía
  - Hipergeneralización
  - Neutralización
  - Asociación Cruzada

Por otro lado, Fernández López (1997) se separa de esta clasificación más cercana a las primitivas de James, Dulay y otros, o Richards, retomando más la línea de Selinker y un enfoque basado en estrategias de aprendizaje y comunicación. De hecho, comenta que es posible hablar de dos grandes ejes o estrategias: la simplificación y la generalización. Aún así, detalla más su clasificación y muestra ocho tipos de explicaciones posibles de los errores de aprendientes:

- Interferencia
- Mecanismos intralinguales: hipergeneralización o generalización
- Analogía con estructuras próximas

[95] Para el caso del francés o el inglés, encontramos taxonomías muy sistematizadas que a menudo sólo cubren el criterio lingüístico y descriptivo. La mayoría de estos trabajos provienen también del análisis computacional de errores (CAE), con lo cual se tiende a la simplificación en la taxonomía para un mejor funcionamiento del editor, que se basa en otras tecnologías como etiquetadores automáticos, analizadores, etcétera. Los trabajos de ELE, muchos de ellos análisis de errores manuales y con un marcado carácter pedagógico, sí han insistido en la causa del error, y guardan además una simetría mayor en los conceptos expresados.

- Influencia de la forma fuerte
- Hipercorrección
- Neutralización de las oposiciones de la lengua meta
- Reestructuración (evasión o abandono de la comunicación, entre otras)
- Formación de una hipótesis idiosincrásica

Para nuestro proyecto, hemos creado una clasificación que trata de aunar los elementos coincidentes de las taxonomías estudiadas. Así, establecemos una taxonomía para el criterio etiológico que se compone de los siguientes errores:

- **Errores interlinguales (<INTER>):** Errores que se producen por la interferencia de estructuras o patrones de otra lengua conocida por el hablante. Pueden distinguirse en:
  - **Errores por interferencia de la lengua materna (<IFL>).**
  - **Errores por interferencia de una tercera lengua (<IFL3).**

- **Errores intralinguales (<INTRA>):** Aquellos que se producen por la interferencia de reglas, estructuras y otros conocimientos ya adquiridos de la lengua meta, porque el hablante trata de salvar ciertos obstáculos de comunicación, utilizando elementos similares de la L2 que ya conoce. Se pueden subdividir, a su vez, en distintos tipos de procesos que causan el error:
  - **Simplificación (<SIM>):** Procedimiento por el cual el aprendiente comete un error al utilizar elementos sintácticos de menor complejidad o más reducido, ya que al estar plenamente adquiridos, el hablante tiene una acceso más rápido a su utilización.
  - **Asociación Cruzada (<ASC>):** Error que proviene al mezclar en la misma estructura normas o elementos de dos formas o categorías distintas. En un principio, ocurre por la similitud de las formas que se entrecruzan.

- **Analogía (<ANA>):** Errores que se producen al comparar con estructuras similares pero que, en verdad, no pueden ser utilizadas de la misma manera.
- **Neutralización (<NEU>):** Cuando se usa la misma estructura para varios elementos que tienen el mismo referente semántico. En muchos casos, se confunde con la simplificación, porque al fin y al cabo, se trata también de una reducción de la complejidad del sistema.
- **Ignorancia de las restricciones de las reglas (<IGN>):** Consiste en aplicar las normas o reglas lingüísticas que ya conoce, pero sin tener en cuenta alguna excepción o caso especial de uso que no permite hacerlo en ese contexto.
- **Aplicación Incompleta de las reglas (<AIR>):** El error se produce al no aplicar correctamente y de forma completa las reglas de una determinada estructura que ya conoce.
- **Hipergeneralización (<HIG>):** Se produce cuando el aprendiente aplica reglas ya conocidas para elementos similares sin tener en cuenta las irregularidades, *generalizando* la regla en momentos en los que no es posible aplicarla.
- **Hipercorrección (<HIC>):** Consiste en un error derivado de la aplicación de excepciones a la regla de forma generalizada para formas que no la necesitan. El aprendiente convierte así en formas irregulares aquellas que no lo son.

- **Errores desconocidos o ambiguos (<UKN>):** etiquetamos bajo esta categoría a todos aquellos errores cuya causa u origen desconocemos, bien porque es ambigua y por tanto susceptible de incluirse en varias categorías, o bien porque no somos capaces de averiguar el proceso que ha podido seguir el aprendiente en la elaboración de su producción en L2.

El criterio etiológico es el más rico a la hora de mostrar los mecanismos que intervienen en la adquisición de una nueva lengua. Sin embargo, es quizá el análisis más difícil de llevar a cabo, ya que, en ocasiones, no es fácil dilucidar cuál es el procesamiento que pone en marcha el aprendiente a la hora de producir su enunciado. Representa el análisis que más información aporta al investigador, sin embargo, también suele ser el más subjetivo.

Cuando no logramos identificar con claridad el proceso de creación de un error, hemos optado por señalarlo bajo la categoría de *errores desconocidos o ambiguos*, ya que preferimos no arriesgarnos a producir una descripción inexacta, que pueda alterar la validez de nuestros resultados.

Otro problema frecuente alude a la posibilidad de que un mismo error comparta más de una categorización, ya que existen errores que implican no sólo interferencias de la lengua materna, sino también mecanismos de manipulación del conocimiento basados en el desarrollo de la interlengua. En estos casos, optamos siempre por marcarlos como *errores ambiguos*, ya que los datos de que disponemos procedentes de la interacción oral, pueden ser insuficientes para permitirnos determinar qué categoría es la más apropiada.

Evidentemente, no podemos olvidar que este análisis tiene un matiz marcadamente subjetivo, ya que el conocimiento que posee el investigador, la norma o variedad con la que compara el enunciado y, por supuesto, su interpretación personal de la forma producida por el aprendiente, darán lugar a variaciones en el análisis, sobre todo, en aquellos errores que hemos denominado como ambiguos. Aún así, intentamos en todo momento que nuestro análisis sea riguroso y objetivo, sometiendo cada error a los mismos puntos de vista, y realizando un análisis en profundidad, como ente particular y también como parte de un conjunto, de forma a asegurar la validez de los resultados y destapar tendencias que puedan ser exploradas por nuevos corpus en el futuro.

7. Conclusiones

En el presente capítulo hemos ofrecido una panorámica de los diferentes tipos de análisis de datos que pueden realizarse con muestras

de lengua. De todos ellos, hemos hecho especial hincapié en el Análisis de Errores, por ser la herramienta central de nuestra investigación.

Así, hemos trazado una breve historia del AE, que tiene en los años 70 su punto más álgido de aplicación, y que sucesivamente, y debido a los cambios producidos en los enfoques didácticos y en las teorías lingüísticas asociadas, ha ido complementándose con otros enfoques como el del Análisis de Actuación, y más recientemente, el Análisis del Discurso.

Pese a las limitaciones que se le achacan, fundamentalmente una falta de homogeneidad en la categorización de errores y una falta de sistematicidad en su aplicación, el AE resulta ser una de las mejores opciones para comenzar un análisis de la interlengua, que, para un mayor rigor, tendría que ser completado con los nuevos enfoques, que contemplan no sólo las dificultades del aprendiente, sino aspectos positivos del uso de la interlengua en distintas situaciones comunicativas.

A lo largo de este capítulo hemos descrito también las características principales del CEA, o Análisis de Errores asistido por ordenador, sucesor natural del AE, y enfoque que pretende mejorar las limitaciones que se atribuyen, con ayuda de grandes colecciones de datos digitalizadas y herramientas y aplicaciones informáticas como editores o etiquetadores automáticos de errores, que pueden además, proporcionar una retroalimentación más completa al aprendiente.

El único inconveniente que encontramos en el CEA es su escasa aplicación por el momento y el número reducido de sus herramientas, que, por otra parte, todavía no resultan del todo efectivas a la hora de tratar la lengua del aprendiente, debido a su complejidad.

Además, hemos disertado sobre las distintas concepciones del error, sus características, y hemos analizado varias consideraciones necesarias para aplicar el AE con rigor.

Finalmente, y siguiendo la metodología de Corder, hemos intentado describir todos los criterios a partir de los cuales puede observarse un error, ahondando principalmente en tres: el criterio lingüístico, el descriptivo y el etiológico. En relación con todos ellos, hemos presentado las taxonomías de utilización más frecuente en los proyectos del ámbito de la enseñanza de lenguas, disciplina que ha favorecido mucho su aplicación. Hemos expuesto sus características y sobre todo, la información sobre la interlengua que nos proporciona.

Hemos de destacar que de todos ellos, el más provechoso y el que aporta más información será siempre el análisis etiológico, puesto que atañe a la causa del error y muestra parte de los procesos internos del aprendiente que influyen en la adquisición de la lengua.

Este recorrido por las distintas modalidades de análisis de errores nos ha permitido, finalmente, presentar las taxonomías para cada criterio de observación que hemos aplicado en el estudio de errores de nuestro corpus de aprendientes, cuyos resultados analizaremos en la segunda parte de la presente tesis.

PARTE SEGUNDA

LA APLICACIÓN

1. ORIGEN, CONCEPCIÓN Y DISEÑO DEL CORPUS DE APRENDIENTES CORAF

1. Introducción

El corpus CORAF (*Corpus ORal de Aprendientes de Francés*) es, como su propio nombre indica, un corpus hablado de aprendientes hispanófonos de francés como lengua extranjera (FLE) en contexto educativo o guiado el cual, en este caso, consiste en tres Escuelas Oficiales de Idiomas (EOI) y la Facultad de Letras de la Universidad de Castilla-La Mancha.

El proyecto CORAF es el objeto central de nuestro proyecto de tesis. Nace con una finalidad pedagógica que tiene los siguientes propósitos:

- Contribuir a ampliar el conocimiento sobre la interlengua de los aprendientes hispanohablantes de FLE;
- Profundizar en la comprensión de cómo se desarrollan las capacidades de interacción oral en los aprendientes, al ser la competencia más difícil de adquirir;
- Realizar un análisis de errores sistemático desde tres puntos de vista: lingüístico, descriptivo y etiológico;
- Poner a disposición de investigadores y docentes unos resultados de investigación susceptibles de mejorar la metodología de enseñanza del francés como lengua extranjera, paliando los problemas más frecuentes de adquisición del oral, y fomentando modalidades de enseñanza adaptadas a las necesidades específicas de los aprendientes.

En el presente capítulo, enumeraremos en primer lugar las motivaciones que subyacen a la decisión de crear este corpus. Posteriormente, realizaremos una primera aproximación al diseño de sus características generales y aspectos técnicos más reseñables, revelando sus componentes y contenido, y explicando su utilidad para distintos propósitos. Finalmente, presentaremos el corpus implementado, comentando algunos de sus datos más reseñables.

2. Motivaciones para la realización del corpus CORAF: ¿Por qué un corpus de lengua *oral* de aprendientes?

En una sociedad en la que la palabra mítica es *globalización*, cada día se hace más necesario el dominio de lenguas extranjeras y, especialmente, el de la comunicación oral. Debido a ello, en el contexto de la enseñanza de las lenguas, las actividades pedagógicas orientadas al desarrollo de la expresión oral ocupan un lugar cada día más importante en el aula de idiomas.

Que el aprendiente sepa comunicarse eficazmente con su interlocutor, no sólo implica que adquiera saberes lingüísticos básicos, sino también que se muestre capaz de adaptarse a diversos entornos de comunicación. Todo ello pasa por el dominio de la dimensión pragmática y discursiva de la comunicación, así como la compresión del contexto cultural en la que esta se produce.

El MCER, -que es el *Marco Común Europeo de Referencia para las Lenguas* propuesto por el Consejo de Europa, para servir de patrón internacional de medición del nivel de comprensión y expresión orales y escritas de las lenguas-, afirma que su enfoque "se centra en la acción en la medida que considera a los usuarios y alumnos que aprenden una lengua principalmente como agentes sociales, es decir, como miembros de una sociedad que tiene tareas (no sólo relacionadas con la lengua) que llevar a cabo en una serie determinada de circunstancias, en un entorno específico y dentro de un campo de acción concreto" (MCER, 2002:20).

En coherencia con ese planteamiento, el aprendiente no adquiere una nueva lengua con un fin meramente académico, sino como agente activo de una sociedad en la que las relaciones multilingües desempeñan un papel económico-cultural fundamental. Eso obligaría, en principio, a los investigadores a centrarse en las competencias de interacción oral de los hablantes como base para el acceso a otras capacidades cognitivas, y como herramienta fundamental para su formación como actor social, capaz de desenvolverse eficazmente en tareas propias de contextos sociales y profesionales variados y cambiantes.

Sin embargo, lo que parece una evidencia, ha sido uno de los aspectos más desatendidos por los expertos. La complejidad del análisis de los procesos de desarrollo de la competencia oral, así como la insuficiencia de datos disponibles para su investigación, han dado lugar a

que la producción científica se oriente más hacia el estudio de las competencias del escrito.

El presente estudio pretende contribuir a paliar esas deficiencias. Sus objetivos centrales son identificar las principales trabas que la interacción oral en lengua meta ofrece al aprendiente, e indagar los modos de mejorarla. Una de las opciones más habituales para el estudio del proceso de adquisición del oral es la realización de un análisis contrastivo de la interlengua, en el que el sistema lingüístico intermedio por el que el aprendiente adquiere la lengua meta es caracterizado, para después compararlo con el de hablantes nativos en contextos similares de uso, o con hablantes de L2 del mismo nivel con una lengua materna diferente. Al igual que para un análisis de errores frecuentes, un corpus de aprendientes, oral en este caso, nos dará el soporte necesario para la investigación, por el elevado número de datos que proporciona, y por la facilidad y rapidez de consulta con programas específicos que también permite.

Además, los corpus de lengua oral suponen un recurso único para la exploración del discurso espontáneo, ya que dicha incursión no podría ejecutarse por otros medios.

En resumen, los aspectos esenciales que nos llevan a abordar la implementación de un corpus de lengua oral de aprendientes son los siguientes:

- la dificultad particular que encuentran los aprendientes para adquirir las competencias orales;
- la desatención que sufre la enseñanza de las competencias orales en los enfoques convencionales, debido precisamente a la dificultad de su desarrollo;
- la escasez de estudios científicos de carácter empírico basados en corpus de aprendientes hispanohablantes de FLE y, por tanto, el desconocimiento general de los procesos mentales implicados en la construcción de la interlengua.

3. Diseño del corpus CORAF

3.1 Aspectos generales

Los corpus, cualquiera que sea su tipología, son esenciales para investigar el uso que se hace de la lengua. Por tanto, un valor inexcusable de los corpus es que sean representativos de la variedad de lengua que queremos investigar o reproducir.

La primera tarea antes de implementar un corpus consiste en realizar un diseño que garantice su validez y calidad. Desde luego, no podemos proceder a crear un corpus sin más, recogiendo los datos de cualquier informante y sin contar con unos criterios de selección. Para llevar a cabo el diseño de una recopilación de corpus, y siguiendo la metodología más habitual en las investigaciones realizadas en el ámbito de la Lingüística de Corpus (Cf. O'Keeffe and McCarthy, 2010), debemos empezar respondiendo a las siguientes preguntas:

- ¿Qué tipo de datos integrará nuestro corpus y cuál sería la cantidad mínima necesaria para la consecución de nuestros objetivos?
- ¿Cómo vamos a recoger las muestras de lengua?
- ¿Qué tipo de tratamiento le vamos a dar a los datos en función del objeto de nuestra investigación?
- ¿Qué formato o apariencia final va a tener nuestro corpus?

El corpus CORAF es un corpus monolingüe, específico o de especialidad, que pretende mostrar una variedad de lengua concreta: la interlengua o lengua de los aprendientes de FLE que comparten el español como lengua materna y que realizan su aprendizaje en un contexto educativo o guiado. En resumen, en la concepción del diseño de nuestro corpus, nos aseguramos del cumplimiento de los siguientes objetivos:

- Que sea representativo y equilibrado;
- Que contenga datos suficientes para nuestros objetivos;
- Que su tamaño sea equiparable al de otros corpus de su misma tipología;

- Que su contenido se ajuste a los objetivos de análisis de nuestra investigación;
- Que posea una forma homogénea y estandarizada;
- Que pueda ser reutilizado para otros fines científicos o pedagógicos.

Para que nuestro corpus sea representativo, hay que intentar cubrir toda la variedad de lengua analizable. En nuestro caso, al tratarse de una lengua que no es la lengua materna, la variedad de lengua incluye los distintos niveles d dominio que implica su adquisición. A este respecto, convinimos que el MCER podría otorgarnos una distribución por niveles aceptada y extendida en el contexto académico, por lo que decidimos buscar aprendientes de FLE de los distintos niveles contenidos en el MCER, desde los más básicos (A1 y A2) hasta los más avanzados (C2). Todo ello nos ofrecería un panorámica amplia sobre el desarrollo de la interlengua, garantizándonos la representatividad necesaria.

Otro de las claves que se han de atender a la hora de realizar un corpus, tiene que ver con su tamaño. Se sabe que los corpus especiales, también denominados específicos o de especialidad, no poseen una cantidad muy elevada de palabras. Habitualmente, contienen una media de 30.000 a 50.000 palabras (aunque hay corpus de hasta más de 200.000 palabras, generalmente de textos escritos). Los corpus orales, por su dificultad de elaboración, suelen ser mucho más reducidos que los escritos (Cf. O'Keeffe and McCarthy, 2010: 38, 67). Como venimos comentando, no existe un tamaño de corpus estándar o ideal. Todo depende de para qué se concibe, y cuál es su campo de aplicación. En nuestro caso, el objetivo es obtener un número de palabras similar al de los corpus habituales en la tipología de corpus orales de aprendientes de L2.

Biber (O'Keeffe and McCarthy, 2010: 70), que realizó distintas investigaciones sobre cuestiones de representatividad de los corpus, afirma que los elementos lingüísticos más comunes (pronombres, verbos conjugados en pasado y presente, preposiciones, etcétera) aparecen ya de manera suficientemente profusamente en muestras de alrededor de 1.000 palabras. Si damos por buena esa referencia, aceptamos que en ningún caso nuestra muestra debe ser inferior a esa cantidad, si queremos garantizar la representatividad de los elementos lingüísticos. Los textos

orales, sobre todo en el contexto de interacción suelen tener una longitud muy reducida en relación con los enunciados escritos. De modo que la tarea de recopilación de muestras representativas del oral no es especialmente fácil.

Por otra parte, es conveniente tener en cuenta que muchos de los corpus de aprendientes conllevan una circunstancia que puede afectar sensiblemente al cómputo general de palabras: en la mayoría de ellos, el número de palabras del/de los entrevistador/es o investigador/es se incluye en la suma total. En el caso de CORAF, cuyo objetivo prioritario es observar la interlengua de los aprendientes, se han aislado del cómputo total de palabras, los enunciados producidos exclusivamente por los aprendientes, ofreciendo no sólo el total de palabras, sino también el número de las producidas exclusivamente por los aprendientes. Nuestra primera idea fue de compilar un corpus de un mínimo de 30.000 palabras de aprendientes, un mínimo de 1.000 palabras para cada muestra individual, distribuidas de forma equilibrada entre los seis niveles representados en el MCER. El resultado ha sido de 33.915 palabras. De modo que nuestro corpus ofrece las garantías necesarias de representatividad, equilibrio, suficiencia de número de datos y ajuste al objeto final de la investigación.

Atendiendo a la variedad de lengua que queremos analizar, es decir la interlengua de aprendientes hispanófonos de FLE, y que realizan su proceso de adquisición en un contexto educativo o guiado, no debemos de perder de vista la especificidad de la variedad de lengua. No es difícil localizar en nuestro entorno sujetos pertenecientes a dicha variedad susceptibles de participar en el estudio, si bien, como veremos más adelante, no son tan accesibles como se podría pensar.

Una vez resuelto lo referente al tamaño, debemos decidir qué tipo de muestras de lengua recoger y con qué tipo de participantes, con el fin de cumplir con los objetivos de representatividad de la variedad elegida. En este punto, y dadas las características de los aprendientes, optamos por entrevistar y grabar a alumnos adultos de la modalidad de enseñanza presencial pertenecientes a todos los niveles del MCER, pues era el público que nos resultaba más accesible para realizar las grabaciones y que requería de menos tramitación de aspectos legales.

Una vez detallados los aspectos básicos relativos al contenido, es importante cumplir las exigencias técnicas relacionadas con el formato. Nuestro proyecto debía cumplir dos premisas básicas: utilizar una

tecnología lo más estandarizada posible, de modo a asegurar la interoperabilidad con sistemas, herramientas y otras aplicaciones que pudieran servir para su análisis o manejo, así como la reutilización del producto para investigaciones posteriores.

La forma textual más aceptada por los ficheros de texto, y la más recomendada por la mayoría de los teóricos de la LC, es el formato texto plano (txt), así como, últimamente, el formato XML, pues garantizan la interoperabilidad con herramientas de todo tipo implementadas en el marco de la Lingüística Computacional.

Para el sonido, utilizamos preferente el formato wav, puesto que guarda una mayor calidad del audio y es fácilmente convertible a formatos más usuales y de menor tamaño como el mp3.

En resumen, componentes básicos del diseño y concepción del corpus CORAF son los siguientes:

Nº DE PALABRAS APROXIMADO	= /> 30.000 palabras (de aprendientes)
CANTIDAD DE PALABRAS POR MUESTRA	=/>1.000 palabras
PARTICIPANTES	Aprendientes adultos de FLE de todos los niveles del MCER en contexto educativo y de lengua materna española
CONTEXTO	Escuelas Oficiales de Idiomas y Universidad
FORMATO PARA LA RECOGIDA DE DATOS	Entrevistas semidirigidas
FORMATO TÉCNICO	Transcripciones en formato txt /xml Archivos de sonido en wav

Tabla 5: Resumen de los aspectos significativos relativos al diseño del corpus CORAF.

3.2 Recogida de datos: elección del contexto de grabación.

La elección del entorno de grabación, y por consiguiente del tipo de aprendientes que van a formar parte de nuestro corpus, responde, ante todo, a la necesidad de crear un corpus representativo y equilibrado. La situación de la recogida de datos en un contexto concreto es determinante para la configuración del corpus y condiciona la selección de los participantes, y por tanto, de su interlengua.

Como ya hemos comentado, queremos mostrar todas las variedades de interlengua del aprendiente de FLE. Ello nos lleva a buscar a participantes de todos los niveles de adquisición de la lengua meta. En la práctica, cada aprendiente tendrá asignado un nivel y unas especificidades propias inmutables. En realidad, los niveles de dominio no han sido atribuidos por individuos, sino en base al nivel supuesto del grupo en la institución.

Evidentemente, en aras de posibilitar una comparación posterior de nuestros resultados con los de otros estudios equivalentes, necesitábamos utilizar niveles de dominio de lengua estandarizados. Desde su creación, el *Marco Común Europeo de Referencia para las Lenguas* (MCER) está siendo utilizado para la elaboración de programas curriculares y métodos de enseñanza de lenguas en toda Europa, por lo que nos garantizaba las cotas de estandarización deseadas a la hora de definir y evaluar los niveles de dominio de una lengua en proceso de aprendizaje. Aunque el modelo no está exento de críticas, era la elección óptima para categorizar a los aprendientes que conforman nuestro corpus, precisamente por ser la referencia común más extendida.

Decíamos que queríamos garantizar a nuestro corpus un volumen suficiente de datos, y que recoja de forma equilibrada todos los niveles establecidos por el MCER. Necesitamos, por tanto, un conjunto de 30.000 palabras, lo que supone realizar entrevistas a un número suficiente de aprendientes para alcanzar ese objetivo.

Cualquier proyecto debe ceñirse a un tiempo específico programado de antemano. Transcribir los datos supone un esfuerzo importante para un investigador y ocupa necesariamente gran parte de su tiempo disponible. Por lo tanto, convinimos en buscar un lugar donde

pudieran realizarse el mayor número de grabaciones de aprendientes de todos los niveles en el menor tiempo posible. De esta forma nació la idea de concentrar nuestro estudio en dos entidades académicas que reunían estas condiciones: las Escuelas Oficiales de Idiomas y la universidad.

Ambos contextos nos permitían llegar a un gran número de aprendientes de todos los niveles y posibilitaban realizar las grabaciones un tiempo y espacio concentrado. Además, nos permitía contar con aprendientes adultos, en su mayoría, y de variadas características socioculturales, personales, motivaciones y estilos de aprendizaje, lo que estaba llamado a enriquecer, sin duda, nuestro corpus.

Por otra parte, en ambos casos es importante señalar que nos daba la oportunidad de acceder a aprendientes considerados como *no cautivos* (Courtillon, 2003), es decir, que cursan la asignatura de francés de forma optativa, por una decisión personal de ampliar sus conocimientos, lo que implica una motivación y una percepción positiva hacia la lengua de estudio. Esta circunstancia nos pareció significativamente interesante, ya que al haber elegido la lengua de estudio de forma voluntaria, creímos que eran susceptibles de involucrarse más fácilmente en la tarea que les propondríamos: la participación en nuestro corpus de lengua oral.

Finalmente, una vez acotado el contexto de grabación, decidimos, por cuestiones logísticas, que el ámbito de búsqueda de estas escuelas de idiomas y universidades sería la región de Castilla-La Mancha. Así, se procedió a contactar por diferentes vías con casi todas las Escuelas Oficiales de Idiomas existentes y con la Facultad de Letras de la Universidad de Castilla-La Mancha (en adelante, UCLM), única entidad educativa en la región que acoge los estudios de Filología Francesa.

3.3 Elección de los aprendientes participantes

Nuestra propuesta de corpus de aprendientes incluye a hablantes de francés como L2 de todos los niveles expuestos en el MCER. Por tanto, se han realizado grabaciones de los aprendientes que se encuentren inscritos en los siguientes niveles:

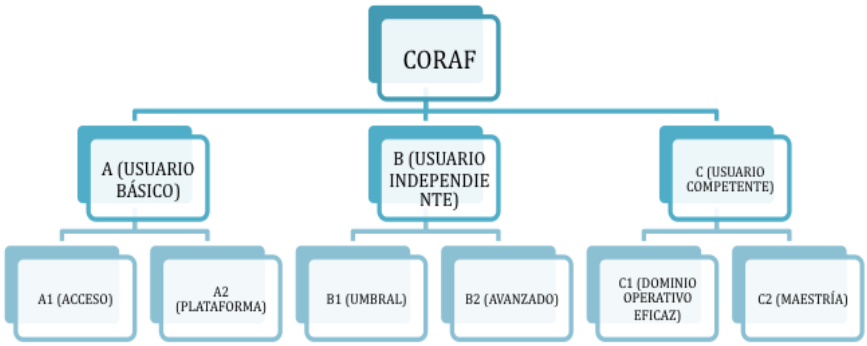

Gráfico 4: Esquema de los niveles de referencia expresados en el MCER (2002) y representados en el corpus CORAF.

Dichos niveles de referencia incluyen una serie de descriptores básicos que recogen todas las competencias comunicativas y generales que ha de adquirir el aprendiente para alcanzar cada uno de ellos. Las competencias generales no se relacionan directamente con la lengua, pero se puede recurrir a ellas para acciones de todo tipo. Las competencias comunicativas sirven al aprendiente para realizar actos por medio de la lengua (MCER, 2002: 20).

En el MCER, dichas competencias se apoyan en un enfoque orientado a la acción, en el que, como ya hemos explicado, el aprendiente adquiere la lengua no con un objetivo estrictamente cognitivo, sino para objetivos de tipo social, cultural o profesional. El uso de la lengua de comunicación consiste en utilizarla "en distintos contextos y bajo distintas condiciones y restricciones, con el fin de realizar actividades que conllevan procesos para producir y recibir textos relacionados con temas en ámbitos específicos, poniendo en juego las estrategias que parecen más apropiadas para llevar a cabo las tareas que han de realizar" (MCER, 2002: 20).

Este enfoque supone que el aprendiente, no se limita a ser un potador de la lengua, sino que se espera de él que y adquiera un rol social en contextos muy variados y cambiantes, en los que son necesarios otro tipo de saberes además del dominio lingüístico, como los de reflexión

sobre el propio aprendizaje o competencias pragmáticas, sociolingüísticas o funcionales. Podemos resumir el bagaje de conocimientos que se espera desarrolle el aprendiente para afrontar eficazmente tareas comunicativas de la siguiente manera:

Gráfico 5: Esquema de las competencias que intervienen en el aprendiente en un enfoque orientado a la acción según el MCER (Fuente: elaboración propia)

Todas estas competencias son las que reproduce la escala global de niveles comunes de referencia propuesta por el MCER (2002: 36), que marcará la descripción general y básica de nuestro corpus de aprendientes[96]:

[96] Evidentemente, el MCER recoge otros muchos indicadores más específicos distribuidos entre las destrezas evaluadas y orientadas a consecución de la expresión oral y escrita y a la comprensión oral y escrita.

USUARIO BÁSICO	A1	Es capaz de comprender y utilizar expresiones cotidianas de uso muy frecuente así como frases sencillas destinadas a satisfacer necesidades de tipo inmediato. Puede presentarse a sí mismo y a otros, pedir y dar información personal básica sobre su domicilio, sus pertenencias y las personas que conoce. Puede relacionarse de forma elemental siempre que su interlocutor hable despacio y con claridad y esté dispuesto a cooperar.
	A2	Es capaz de comprender frases y expresiones de uso frecuente relacionadas con áreas de experiencia que le son especialmente relevantes (información básica sobre sí mismo y su familia, compras, lugares de interés, ocupaciones, etc.) Sabe comunicarse a la hora de llevar a cabo tareas simples y cotidianas que no requieran más que intercambios sencillos y directos de información sobre cuestiones que le son conocidas o habituales. Sabe describir en términos sencillos aspectos de su pasado y su entorno así como cuestiones relacionadas con sus necesidades inmediatas.

USUARIO INDEPENDIENTE	B1	Es capaz de comprender los puntos principales de textos claros y en lengua estándar si tratan sobre cuestiones que le son conocidas, ya sea en situaciones de trabajo, de estudio o de ocio. Sabe desenvolverse en la mayor parte de las situaciones que pueden surgir durante un viaje por zonas donde se utiliza la lengua. Es capaz de producir textos sencillos y coherentes sobre temas que le son familiares o en los que tiene un interés personal. Puede describir experiencias, acontecimientos, deseos y aspiraciones, así como justificar brevemente sus opiniones o explicar sus planes.
	B2	Es capaz de entender las ideas principales de textos complejos que traten de temas tanto concretos como abstractos, incluso si son de carácter técnico siempre que estén dentro de su campo de especialización. Puede relacionarse con hablantes nativos con un grado suficiente de fluidez y naturalidad de modo que la comunicación se realice sin esfuerzo por parte de ninguno de los interlocutores. Puede producir textos claros y detallados sobre temas diversos así como defender un punto de vista sobre temas generales indicando los pros y los contras de las distintas opciones.

USUARIO COMPETENTE	C1	Es capaz de comprender una amplia variedad de textos extensos y con cierto nivel de exigencia, así como reconocer en ellos sentidos implícitos. Sabe expresarse de forma fluida y espontánea sin muestras muy evidentes de esfuerzo para encontrar la expresión adecuada. Puede hacer un uso flexible y efectivo del idioma para fines sociales, académicos y profesionales. Puede producir textos claros, bien estructurados y detallados sobre temas de cierta complejidad, mostrando un uso correcto de los mecanismos de organización, articulación y cohesión del texto.
	C2	Es capaz de comprender con facilidad prácticamente todo lo que oye o lee. Sabe reconstruir la información y los argumentos procedentes de diversas fuentes, ya sean en lengua hablada o escrita, y presentarlos de manera coherente y resumida. Puede expresarse espontáneamente, con gran fluidez y con un grado de precisión que le permite diferenciar pequeños matices de significado incluso en situaciones de mayor complejidad.

Tabla 6: Cuadro resumen de la escala global de los niveles comunes de referencia. (Fuente: MCER, 2002: 36)

En principio, decidimos realizar grabaciones con el mayor número posible de aprendientes voluntarios que pudiésemos encontrar. Desde luego, este requerimiento no garantizaba de antemano el número de participantes que exigía nuestro estudio, pero consideramos que el hecho de que los aprendientes no fueran obligados a realizar la entrevista podría beneficiar al proyecto. Entre otras cosas porque la disposición positiva para hablar eliminaba gran parte de la presión externa, lo que favorecería, sin duda, la calidad de la producción oral.

Es innegable que las propias características de la interacción conllevaban una cierta ansiedad (situación comunicativa no habitual, ante

un entrevistador desconocido, sin conocer la tarea previamente y, además, estando delante la grabadora digital). Pero pensamos que con un público voluntario podríamos paliar otros factores externos importantes de tensión (por ejemplo, obligación por parte del profesor, sentimiento de necesidad de alcanzar los objetivos propuestos por este y de dar la imagen que se espera de él como aprendiente de un determinado nivel, etcétera) que previsiblemente sobrecargarían la interacción de un exceso de tensión.

No obstante, al diseñar las grabaciones, nos fijamos la consecución, en un primer momento, de al menos 10 entrevistas de aprendientes por nivel. El desarrollo de las grabaciones y la dificultad de conseguir hablantes para ciertos niveles concretos hizo, como explicaremos más adelante, que hubiera que reconsiderar el diseño, sobre todo con vistas a la posterior transcripción y análisis de errores, ya que el corpus quedaba desequilibrado, al no encontrar el mismo número de locutores para todos los niveles (concretamente, existe un desequilibrio para el nivel de usuario independiente: B1 y B2, con 6 y 7 hablantes entrevistados).

Por otro lado, el hecho de realizar las grabaciones en la EOI y en la Facultad de Letras, como ya hemos comentado anteriormente, nos facilita un acceso a un público amplio con unas características muy distintas entre sí en cuanto a edad, nivel de estudios, profesión u origen geográfico. Todos estos datos que, como veremos, están convenientemente reflejados en la cabecera de nuestras transcripciones, resultan muy interesantes para análisis sociolingüísticos posteriores.

Así, las características básicas que debían ostentar los aprendientes que formarán parte del corpus eran simplemente pertenecer a la entidad académica, estar matriculados en un nivel determinado y haber recibido un tiempo de instrucción a lo largo de su proceso de aprendizaje, realizar la entrevista en su totalidad y haber firmado la autorización para el tratamiento y difusión de los datos.

4. Metodología de recogida de datos

4.1 Consideraciones generales

El proceso de recogida de datos o grabación es, sin duda, junto con el diseño, la base para el buen desarrollo de un corpus. Así, esta fase viene determinada sobre todo por el tipo de análisis previsto de los datos (u objetivo final) y por los elementos de lengua que queramos incluir en el corpus, entendiendo que respeten las características necesarias de representatividad y equilibrio.

A la hora de crear y diseñar un corpus hay que tener en cuenta algunos factores como el tipo o la variedad de lengua que queremos abarcar, por medio de qué tareas vamos a intentar conseguir esa muestra de variedad de lengua y qué tipo de variables hemos de controlar en el aprendiente.

Para autores como Ellis y Barkhuizen (2005: 21), lo que los aprendientes conocen se refleja de forma más adecuada en la comprensión del *input* y en la lengua que producen. Además, se cree que es en la producción donde se ven los mayores indicios de lo que el aprendiente ha adquirido, por lo que es el método más extendido para la investigación en adquisición de segundas lenguas, ya que privilegia la recogida de muestras de habla o lengua escrita por encima de otras actividades. De hecho, tanto hablar como escribir se consideran actividades lingüísticas naturales.

Existen varias formas de lograr la producción de datos por parte del aprendiente, que se agrupan en dos enfoques específicos (Dulay et al., 1982: 247):

- Tareas de comunicación espontánea
- Tareas de manipulación lingüística

Las primeras se centran en la comunicación de ideas u opiniones en vez de en formas lingüísticas y se basan en la creencia, como también expresaban Ellis y Barkhuizen, de que el aprendiente utilizará de forma inconsciente reglas gramaticales y otros elementos lingüísticos que ha adquirido y que pueden adaptarse a la situación comunicativa en la que se encuentra.

Las tareas de manipulación lingüística requieren del aprendiente la producción consciente de ciertos elementos lingüísticos concretos, llevándoles a manipular y cambiar oraciones o partes de ella, resultando interesantes para dilucidar aspectos de la conciencia metalingüística.

Es evidente que nuestra idea de recoger una producción espontánea choca con el uso de tareas de manipulación lingüística, ya que se trata de tareas no comunicativas en las que se suele utilizar una lengua muy artificial, que se centra en formas muy estereotipadas y aisladas.

Para nuestro estudio resulta más conveniente una tarea de comunicación espontánea, porque se articula sobre una lengua natural y puede reflejar de forma más clara la interlengua del aprendiente. No obstante, hemos de saber que el habla espontánea posee algún inconveniente, puesto que es posible que en el transcurso de la conversación no aparezcan determinadas estructuras por su dificultad o por su escasa utilización, además de la repetición constante de otras ya adquiridas y que proporcionan seguridad al aprendiente durante la interacción.

Muchos investigadores, a la hora de realizar estas grabaciones, eligen distintas tareas o metodologías para conseguir los datos de la interlengua que necesitan. Como señalan Ellis y Barkhuizen *(op. cit*), se distribuyen en las tareas de elicitación clínica y las de elicitación empírica. En la inducción empírica se intenta que el aprendiente incorpore en algún momento de su producción aspectos que son de interés para el investigador. Por el contrario, en la inducción clínica, se deja que el aprendiente tenga una producción libre, con datos de todo tipo, primando las tareas que suponen significativamente algo para el aprendiente.

Así, como venimos subrayando, nuestro objetivo es definir la lengua del aprendiente de FLE en interacción oral, por lo que creímos necesario utilizar un enfoque clínico y establecer una conversación informal con los aprendientes, ya que, como ya hemos comentado, es en este tipo de comunicación/tarea en la que el aprendiente se encuentra con más libertad y puede elaborar su discurso según sus conocimientos y los aspectos lingüísticos ya adquiridos.

Para ello, realizamos una de las tareas más frecuentes, las entrevistas orales, con la intención de generar conversaciones

espontáneas de las que obtener muchos datos sobre la producción sobre la interlengua.

4.2 Entrevistas semidirigidas como protocolo de recogida de datos

Las entrevistas orales suelen ser un método habitual de recogida de datos, porque aportan mucha información sobre la interlengua y sobre otros factores internos y externos que influyen en el aprendiente. Generalmente se utilizan en estudios en los que no se busca validar una hipótesis concreta o donde se tiene decidido el tipo de análisis que hay que realizar, sino que se prima la consecución de datos abundantes y variados. Por tanto, es muy utilizada en estudios cualitativos y de observación etnográfica.

Las entrevistas orales consisten en conversaciones informales con el aprendiente. Pueden ser completamente espontáneas o bien, seguir un esquema preparado de antemano, que, en su caso, puede dejar libertad de respuesta al aprendiente (entrevistas semidirigidas), u orientarlo totalmente, no permitiéndole hablar o comentar temas que no sean los contenidos en las preguntas (entrevista dirigida).

Las entrevistas orales en general, como conversaciones informales que son, podían, en nuestro caso, dar lugar a multitud de temas y formas lingüísticas distintas para cada uno de los aprendientes, por lo que la comparación o el análisis entre aquellos del mismo nivel resultaría prácticamente inabarcable. Para favorecer la comparación y la aparición de estructuras similares en todos ellos optamos por desarrollar un esquema de entrevista que guiara todas nuestras grabaciones. Es lo que hemos denominado anteriormente una entrevista semidirigida, ya que tiene como base un guión general de preguntas para todas las entrevistas, pero deja que el aprendiente se exprese libremente en sus respuestas, dando lugar incluso a nuevas preguntas relacionadas con lo expuesto por él, lo que le aporta un mayor matiz de espontaneidad. De esta manera, por un lado, nos garantizábamos una coherencia y la posibilidad de comparación de los datos, y por otro, ayudábamos al aprendiente en el desarrollo de la conversación, sobre todo en aquellos niveles más básicos donde todavía no tienen una autonomía en su expresión oral.

Hemos de señalar que el hecho de realizar una entrevista semidirigida no resta, como podría pensarse, espontaneidad a la situación comunicativa. Es cierto que vamos a controlar, de alguna forma, la entrevista y algunos de los temas que van a aparecer (y por lo tanto, el léxico y las estructuras lingüísticas asociadas), pero sigue resultando espontánea al no conocer los aprendientes el esquema de preguntas con anterioridad y al dejar que se expresen libremente sobre ellas.

Además, las entrevistas mantienen el esquema general de preguntas, pero en la mayoría de ellas, existen nuevas preguntas o aclaraciones que surgen de las propias respuestas de los aprendientes, con lo que podríamos afirmar que se aproxima a lo que consideramos una conversación informal habitual. La mayoría de los corpus de aprendientes que conocemos, además de ser corpus escritos, se centran sobre todo en tareas de producción de diferentes géneros, o en tareas predeterminadas que orientan la producción del aprendiente hacia un tipo determinado de discurso, por lo que CORAF introduce novedades importantes con respecto al resto de corpus, pues abarca varios de discursos como la descripción, la argumentación o la narración.

4.3 Concepción y diseño de las entrevistas semidirigidas

Las entrevistas pretenden mostrar los aspectos ya adquiridos de la lengua meta y hacer sobresalir aquellos otros que suponen un escollo para el aprendizaje o que pueden inducir a error. Por lo tanto, era necesario basarnos en los conocimientos de cada uno de los niveles a la hora de construir nuestra guía de preguntas.

Así, cada nivel posee un esquema de preguntas adaptado. Para ello, hemos tenido en cuenta el programa curricular que se establece para cada nivel en las Escuelas Oficiales de Idiomas a las que hemos tenido acceso y lo definido por el MCER para cada nivel en lo que a expresión e interacción oral se refiere.

Las entrevistas más sencillas y de menor duración son las de A1 (nivel acceso), aumentando progresivamente en complejidad para el resto de niveles (usuario independiente y usuario competente). Cada uno de ellos incluye no sólo conocimientos desarrollados para esa fase, sino que profundiza en conocimientos previstos para niveles anteriores. Es decir,

cada nivel recoge, de alguna manera, parte de las preguntas de los precedentes, pudiendo así tener la posibilidad de descubrir si dichas competencias han sido adquiridas o si existen un dominio deficiente o errores *fosilizados*.

Para lograrlo, contamos con preguntas que versan sobre aspectos de la vida cotidiana del entrevistado en ámbitos como la vida social, cultural o profesional. Si nos remitimos al MCER, sabemos que la enseñanza debe reflejar usos relevantes de la lengua, tanto para ámbitos personales (o públicos) como profesionales o educativos (MCER, 2002: 26). El aprendiente es un agente social que ha de adaptarse a todos estos contextos, utilizando para ello las estructuras lingüísticas, pragmáticas y funcionales coherentes con cada situación. Por lo tanto, el realizar entrevistas donde se abarcan este tipo de temas fomenta el uso de todas esas estructuras, y ayuda también al aprendiente a practicar la lengua en los términos en los que se expresa el MCER.

La elección de los temas es un aspecto también relevante, porque evidentemente, como ya hemos comentado, va a marcar las pautas del léxico y las estructuras sintácticas que el aprendiente va a utilizar. En la medida de lo posible, hemos tratado de abarcar temas generales y variados, pero siempre cercanos a la cotidianidad de los aprendientes, como recomienda el MCER (2002: 53). Primero, para reducir su ansiedad ante temas o un léxico menos familiares, lo que les permite hablar más sosegadamente, con más seguridad y por más tiempo, garantizándonos una muestra de habla equilibrada y suficiente. Segundo, porque en el desarrollo de las entrevistas nos basamos en el currículo de las distintas Escuelas Oficiales, por lo que tratamos de insertar el vocabulario que ya deben conocer. Debido a la variedad de diferentes experiencias sociales, profesionales y educativas, los participantes de nuestro corpus hablan de los mismos temas desde múltiples puntos de vista, lo que garantiza también una cierta variedad en la forma de expresión de los elementos tratados.

El MCER subraya en lo relativo a este punto los temas de comunicación que han de centrar los actos comunicativos, incluyendo los siguientes (MCER, 2002: 53):

- Identificación Personal.
- Vivienda, hogar y entorno.

- Vida cotidiana.
- Tiempo libre y ocio (aficiones e intereses, radio y televisión, cine, teatros y conciertos, etcétera).
- Viajes.
- Relaciones con otras personas.
- Salud y cuidado corporal.
- Educación.
- Compras.
- Comidas y bebidas.
- Servicios públicos.
- Lugares.
- Lengua extranjera.
- Condiciones atmosféricas.

De todos ellos, nuestras entrevistas contienen preguntas relativas a la mayoría de los temas expuestos para garantizar una continuidad y un paralelismo con aquello que están aprendiendo.

Asimismo, hemos de destacar que se incluyen algunas preguntas sobre aspectos culturales de la lengua y otras relativas a su experiencia de aprendizaje de la lengua meta. Por un lado, las primeras nos ayudan a valorar el nivel de conocimiento integral de la lengua, puesto que creemos que conocer una lengua incluye también sus componentes culturales. Por otro, las preguntas relativas a su experiencia de aprendizaje de la lengua meta y a la percepción de la misma, nos ayudan a entender las distintas motivaciones que posee el aprendiente, así como otros aspectos que podrían influir positiva o negativamente en el proceso de adquisición. Como hemos indicado en anteriores capítulos, en el aprendiente influyen factores internos y externos, como la motivación, la experiencia en el conocimiento de otras lenguas, el fin por el que acceden a la instrucción, etcétera. Todos ellos influyen de manera importante en el aprendizaje, y por tanto, han de ser estudiados si queremos caracterizar la interlengua del aprendiente en su totalidad.

La entrevista semidirigida creemos que supone una forma muy completa de inducción a la producción del aprendiente, puesto que no sólo conlleva un análisis de su producción oral, sino que nos ofrece datos sobre su competencia o habilidad para:

- Comprender el significado de un mensaje en la lengua meta (en este caso, las intervenciones del entrevistador), elemento que se cree necesario para la adquisición por parte del aprendiente de las formas y la manera de codificar el mensaje en L2;
- Producir y modificar el *input*, lo que supone que sea capaz de desarrollar aspectos concretos relacionados con la morfología y la sintaxis
- Atender a la forma de sus producciones en lengua meta para codificar correctamente el mensaje, lo que le ayuda a desarrollar y ampliar su propio sistema lingüístico

Es decir, que las entrevistas proporcionan no sólo una oportunidad para el aprendiente de hablar en la lengua meta, sino también de escuchar mensajes en esa lengua, mensajes que tiene que decodificar, comprender, e interpretar para responder adecuadamente, atendiendo a la forma, al significado de la L2 y siempre con coherencia y adecuación al contexto de la interacción.

En resumen, las entrevistas, que tuvieron una duración desde los doce minutos para hablantes del nivel A1 y A2, hasta los veinte minutos de los hablantes competentes, C1 y C2, han seguido una estructura similar a la que reflejamos en el siguiente esquema, aunque sensiblemente diferente en cuanto al contenido en función del nivel:

Gráfico 6: Esquema general del contenido de las entrevistas semidirigidas (Fuente: elaboración propia).

En cuanto a la puesta en práctica de la entrevista, se decidió seguir ciertas pautas para el desarrollo de la misma:

- Realizar un primer período de presentación del aprendiente, con preguntas sencillas, que conoce desde los primeros pasos en el entorno educativo, para reducir su ansiedad y lograr que se vaya acomodando a la situación comunicativa;
- Seguir el orden de las preguntas para cada aprendiente del mismo nivel, aunque modificando o ampliando, en su caso, con nuevas preguntas para obtener más información sobre aspectos interesantes de las respuestas de los aprendientes, o bien, para ayudarles en momentos de duda, bloqueo o nerviosismo;
- No corregir los errores producidos por el aprendiente, salvo en momentos en los que lo demande expresamente o tenga dudas sobre expresiones o elementos léxicos, sintácticos, etcétera;
- Interrumpir al aprendiente lo menos posible, de forma que se exprese libremente en todas las preguntas, intentando fomentar la espontaneidad, una conversación natural, donde el participante pueda, llegado el caso, modificar o alterar el curso de la conversación con sus preguntas, sugerencias, ideas, etcétera (aunque siempre ha de tratarse volver al esquema general de entrevista);
- Evitar forzar al aprendiente a hablar si no tiene nada que decir o si se encuentra bloqueado. Tratar de ayudarle a reconducir la situación o a modificar su discurso para continuar la entrevista;
- Atender a las peticiones del aprendiente en caso de duda, de no comprensión de las preguntas del entrevistador o de demanda de ayuda ante alguna dificultad precisa.

Finalmente, hemos de señalar que todas las entrevistas se realizan, como ya hemos indicado, con aprendientes voluntarios, que, además, desconocen la tarea de antemano e incluso el fin de la misma. Los aprendientes han sido informados por sus profesores-tutores de que formarán parte de un estudio sobre la interlengua y sobre el tipo de proyecto que lo enmarca, concediendo una autorización por escrito, pero desconociendo que se analizarán los errores que comete. La intención es que hablen naturalmente, como lo harían en una conversación corriente

con un hablante nativo o no nativo con competencias más avanzadas. Al conocer de antemano el objetivo del estudio, hubieran podido alterar su discurso para evitar los errores o sentirse demasiado presionados en el transcurso de la tarea, dificultándose así la producción oral.

Una vez finalizada la entrevista, se les informó de cuál sería el próximo uso del corpus (el análisis de errores frecuentes), dándoles la oportunidad de continuar o no colaborando con el estudio.

4.4 Aspectos legales

Realizar un corpus, tanto oral como escrito, supone tener en cuenta ciertos aspectos legales relacionados con la publicación y difusión de datos personales. En todos los casos es necesario que los participantes conozcan con toda claridad para qué servirán sus contribuciones y cómo serán difundidas, ya que contienen datos que aluden a su vida privada, lo cual implicaciones legales.

En primer lugar, y teniendo en cuenta el marco legal, optamos por realizar grabaciones de hablantes adultos, de forma que pudieran ser ellos mismos los que nos otorgaran su permiso de grabación, análisis y difusión. En caso de hablantes menores, este permiso debería ser otorgado por los padres, lo que hubiera ralentizado el proceso de grabación[97].

Antes de cada entrevista, se les informó asimismo de que iban a ser grabados y de que formarían parte de un estudio, pudiendo dar o no su consentimiento para realizar la grabación. Es una vez finalizada cuando se les explica el fin de la misma y se les presenta un formulario de permiso de grabación tipo, que deberán leer, completar y firmar en caso de estar de acuerdo.

En el permiso de grabación, los aprendientes conocen los datos del proyecto y autorizan al investigador a llevar a cabo la grabación de la voz, su transcripción, tratamiento, publicación y difusión en estudios relacionados. Se les garantiza siempre el anonimato y la posibilidad de oír

[97] En nuestro caso, realizamos grabaciones de dos hablantes menores, de los que no hemos realizado su transcripción al no contar con el permiso paterno.

de nuevo la grabación y denegar la autorización en un futuro por cualquier motivo.

En lo que respecta a la anonimización de los datos personales de los aprendientes, en la transcripción se han intentado eliminar todos aquellos datos referentes al lugar de residencia y nombre del aprendiente. Así, en la cabecera de las transcripciones se proporcionan datos sociodemográficos, aunque algunos de ellos como la edad o el rango de estudios aparecen determinados por unas escalas representadas por números o letras. En ningún momento aparece el nombre y apellidos de los participantes en la transcripción, siendo representados siempre por sus iniciales.

Por otra parte, el nombre del fichero trata también de ayudar en la anonimización de los participantes estableciendo sólo el nivel de dominio al que pertenecen, si es hombre (Man, M) o mujer (Woman, W), y el número de la entrevista, (por ejemplo, A1W02), de forma que sin conocer el corpus, en ningún momento podemos reconocer de antemano de qué participante se trata.

Finalmente, cabría señalar que estamos estudiando la posibilidad de suprimir también del audio los datos personales significativos, cortando el sonido, o añadiendo algún tipo de ruido sobre dichos segmentos, de forma que el anonimato quede claramente garantizado.

5. Dificultades previas a la compilación del corpus CORAF

En cada etapa de desarrollo de un proyecto de investigación surgen algunas dificultades. Hablaremos aquí con más detalle de las dificultades surgidas durante la fase de implementación, por ser aquella en la que se concentra la mayor cantidad de ellas, pero no queremos pasar por alto algunas de las que surgieron en el punto de partida del estudio.

En primer lugar, conviene destacar que el hecho de no disponer de ningún corpus oral de aprendientes de FLE disponible en nuestra área de conocimiento, en el que poder apoyarnos o inspirarnos, ya supone una pequeña dificultad de partida. Si bien teníamos ideas claras sobre la metodología y el diseño que íbamos a adoptar para nuestro corpus, la consulta de otros estudios equivalentes hubiera constituido una ayuda

valiosa, pues nos hubiésemos inspirado en sus aciertos e hubiéramos intentado no repetir los mismos errores.

En segundo lugar, es importante mencionar un obstáculo, que sin duda, es mucho más significativo y redunda directamente en el contenido de nuestro corpus: la búsqueda de contextos de grabación. Si bien habíamos previsto contar, *a priori*, con al menos trece lugares distintos para llevar a cabo las grabaciones de alumnos de EOI (número total de escuelas para la región de Castilla-La Mancha), nos encontramos al final con una tímida colaboración por parte de ellas, logrando que tan solo tres accedieran a colaborar. En consecuencia, nuestro estudio no pudo contar con un número mayor de participantes y, sobre todo, de lugares distintos, para una más adecuada comparación de la interlengua y una mayor representatividad de la muestra.

Muchas pueden ser las causas para la baja participación por parte de las EOI, y tras haber analizado el problema con los departamentos de aquellas que sí han participado, citaremos las siguientes:

- La necesidad de impartir muchos conocimientos en cada curso en poco tiempo, hacía que no pudieran dedicar un tiempo a actividades fuera de las previstas en la programación;
- Desconocimiento de las investigaciones en Lingüística de Corpus y su utilidad;
- Interacción poco frecuente entre las EOI y el mundo académico universitario e investigador;

Otro de los aspectos que supone una dificultad es el tiempo disponible para las grabaciones, que es bastante limitado. Las EOI realizan tantas actividades que, aún accediendo a realizar las grabaciones y planeándolo de manera que ocasionen las menores molestias posibles para el desarrollo de la instrucción, supone un tiempo extra que no están dispuestas a perder (no ocurre en todos los casos ni para todos los profesores-tutores). El tiempo concedido de permanencia en la EOI realizando las entrevistas es, por tanto, limitado, lo que influye en la calidad de las producciones y en el número final de aprendientes participantes, puesto que disponer de mayor tiempo supone poder entrevistar a un mayor número de personas.

Por último, existe una dificultad asociada al nivel real o ya adquirido del aprendiente. El hecho de basarnos en un criterio externo para seleccionar el nivel de los aprendientes, como es el curso en el que están escolarizados en sus respectivas Escuelas Oficiales de Idiomas o en la Facultad, nos asegura, *a priori*, que los sujetos puedan ser comparables en términos de competencia lingüística. Evidentemente, encontraremos diferencias en su corrección o fluidez, pero al menos, suponemos que han de compartir una base común. También nos aseguramos de que hayan recibido el mismo tipo de instrucción, durante el mismo tiempo (al menos, mientras han estado escolarizados en la entidad educativa que los acoge actualmente) y con las mismas características.

Este aspecto resulta, a su vez, otra de sus limitaciones ya que, una vez analizada la entrevista, resulta que muchos de los aprendientes no habían alcanzado el nivel que se les presupone.

El caso más claro se puede observar con aquellos hablantes de C2. Evidentemente, partimos quizá de una premisa falsa, y es que aquellos aprendientes del nivel más alto de la escuela de idiomas, y de 5º de Filología Francesa, pertenecerían al nivel C2. A todas luces, se trata de una falacia, puesto que el nivel expresado por el MCER para C2 considera al hablante prácticamente bilingüe, algo que, sin duda, no nos encontraremos en ninguno de los aprendientes que forman parte de nuestras entrevistas. Realmente, aquellos errores y características propuestas para C2 podrían asimilarse a un nivel inferior, puesto que, como afirmamos, los aprendientes no tienen aún, siguiendo todos los índices del MCER, el nivel que deberían. En futuras investigaciones, para el nivel C2, tendríamos que buscar alumnos que estuvieran en posesión del título DALF C2 (*Dîplome Approfondi de Langue Française*), ya que se trata del mejor test que existe para probarlo.

Las causas de esta diferenciación en el nivel de los aprendientes pueden ser fundamentalmente tres:

- Primero, una diferencia entre el nivel en expresión escrita y oral: es frecuente encontrar aprendientes que en su competencia escrita poseen el nivel que les corresponde por escolarización, y que en competencia oral todavía no lo han adquirido, situándose en un nivel inmediatamente inferior;
- segundo, el tipo de obtención de datos escogido (*elicitation task*), es decir, una entrevista semidirigida o centrada, que genera una

tensión suplementaria ante la tarea desconocida, que además, es grabada, y la imposibilidad de reflexionar las respuestas por su carácter espontáneo;

- finalmente, el momento escogido para realizar las entrevistas antes de la conclusión del curso académico.

En próximos estudios o en una futura ampliación o revisión de CORAF, quizá fuera necesario realizar un test inicial para valorar el nivel del aprendiente, que, unido al propuesto por el entorno académico, puede ayudarnos a determinar más objetivamente a qué nivel de competencia del MCER pertenece. Asimismo, hemos de señalar que la decisión de no incluir un pre-test en la entrevista oral vino motivada, sin duda, por la falta de tiempo para realizar la grabación, al que hemos aludido anteriormente.

En resumen, debemos destacar que las principales dificultades previas han estado relacionadas con la ausencia de participación de muchas de las EOI con las que se pretendía contactar y el poco tiempo disponible para llevar a cabo las grabaciones, lo que redunda, en ambos casos, en un menor número de participantes disponibles para nuestro corpus.

6. Diseño final

6.1 CORAF: una primera descripción

El corpus CORAF es un corpus oral monolingüe de aprendientes de francés como lengua extranjera y de lengua materna española, que recoge 30 muestras de habla espontánea de 34 participantes (30 aprendientes y 4 entrevistadores). CORAF contiene un total de 61.092 palabras, 33.915 correspondientes a la producción de aprendientes, y una duración de más de siete horas de grabación.

Además, encontramos participantes para cada nivel expuesto en el MCER (en concreto, A1, A2, B1, B2, C1 y C2), inscritos en tres escuelas oficiales de idiomas de la región y en la Facultad de Letras de la Universidad de Castilla-La Mancha (UCLM), aunque no todos originarios de la región. Pese a querer reflejar un equilibrio en cuanto al

género de los participantes, no ha sido posible, contando al final con 18 mujeres y 12 hombres.

Siguiendo la diferenciación establecida por Granger (2004) y Pravec (2002), podemos definir a CORAF como un corpus académico, ya que se concibe en el ámbito de la investigación universitaria, para un proyecto de investigación concreto y posee el tamaño habitual de este tipo de corpus, que se sitúa en torno a las 30.000 palabras.

Además, al ser un corpus con fines pedagógicos, podemos definirlo a su vez como un corpus *for delayed pedagogical use* (DPU) o corpus para uso pedagógico posterior (Granger, 2009), puesto que los aprendientes que forman parte del corpus no lo usan directamente como material de aprendizaje, es decir, en enfoques directos o DDL. Simplemente nace con la intención de servir en un análisis de la interlengua o para elaborar, por ejemplo, materiales y aplicaciones más específicas para aprendientes de características similares a los informantes del corpus[98].

Por tanto, podemos definir a nuestro corpus como de tipo trasversal, puesto que abarca interacciones orales de aprendientes de distintos niveles en un único momento de su proceso de aprendizaje. Aunque para caracterizar la interlengua del aprendiente resultarían mucho más útiles los corpus longitudinales, los problemas de implementación y el escaso tiempo para la recolección de interacciones nos obliga a realizar un corpus de este tipo, que, por otra parte, suele ser el más frecuente hoy en día en el ámbito de la LC.

No obstante, Granger (2004: 131) señala que los corpus que se componen de muestras de un momento concreto pero provenientes de aprendientes de distintos niveles de competencia casi podrían denominarse longitudinales (en sus propias palabras, *quasi-longitudinal corpora*). La misma opinión la recogen Ellis y Barkhuizen (2005:97), que aluden a corpus *pseudo-longitudinal.* Aunque se muestran críticos hacia este tipo de estudios, puesto que deben de darse unas características determinadas de los aprendientes, muy controladas y exactas en todos

[98] No obstante, en futuras ampliaciones del corpus, no se descarta que pueda llegar a ser un corpus de tipo IPU, *inmediate pedagogical use,* o corpus para uso pedagógico inmediato, donde los aprendientes sean a la vez productores y receptores de los datos del corpus, ya que se prevé, con fines de investigación, un uso directo del mismo en un entorno académico.

para que el estudio tenga validez. Sin embargo, se trata de una tipología no muy utilizada, ya que los expertos en adquisición de lenguas los siguen considerando trasversales y prefieren a los realmente longitudinales para el desarrollo de sus estudios. Si seguimos la definición de Granger, CORAF podría ser un corpus casi longitudinal, aunque coincidimos con Ellis y Barkhuizen en su apreciación, por lo que creemos que no podría considerarse fielmente un estudio de dicha tipología.

Para su recolección se utiliza como metodología el desarrollo de entrevistas semidirigidas, descrito anteriormente, cuya duración media final total es de 14 minutos y 45 segundos, ya que aumenta en duración del mismo modo que aumenta el nivel de competencia del aprendiente.

Finalmente, podemos resumir la estructura y la composición de nuestro corpus a partir de la siguiente tabla:

NIVEL MCER	DURACIÓN TOTAL	DURACIÓN MEDIA	Nº ENTREVISTAS	PALABRAS TOTALES	PALABRAS APRENDIENTE
A1	1:00:24	12m 05s	5 (2H/3M)	6.989	2.506
A2	1:05:22	13m04s	5 (3H / 2M)	8.503	4.110
B1	1:14:19	14m52s	5 (1H / 4M)	9.699	4.908
B2	1:19:46	15m57s	5 (2H / 3M)	11.279	6.858
C1	1:20:28	16m06s	5 (2H / 3M)	12.365	7.867
C2	1:22:04	16m25s	5 (2H / 3M)	12.257	7.666
TOTAL	**7:22:23**	**14m45s**	**30 (12H/ 18M)**	**61.092**	**33.915**

Tabla 7: Tabla resumen de las características básicas del corpus CORAF.

Nuestro corpus se interesa demás por las características que puedan influir sobre su aprendizaje, como el conocimiento que posean de otras lenguas o si han tenido experiencias de inmersión lingüística, en general, estancias prolongadas en países francófonos. Así, 26 participantes poseen algún conocimiento de otras lenguas (sobre todo, de inglés), lo que supone un 86%, y sólo 4 no tienen formación o no constan datos sobre ellos. De los 26 participantes con conocimiento de otras lenguas, 7 han recibido formación de dos lenguas (generalmente inglés e italiano) y 2 aprendientes de más de dos lenguas: uno que conoce el inglés, el alemán y el italiano, y otro participante que conoce además algunas lenguas menos frecuentes como el árabe o el chino.

Finalmente, un 23% de los participantes ha realizado también estancias prolongadas en países de habla francófona, la mayoría de ellos durante sus estudios universitarios gracias a los programas de intercambio Erasmus, otros por motivos de trabajo o realizando cursos de idiomas, intercambios o estancias en familias durante el verano. Casi todos los participantes han visitado y conocen algunas ciudades importantes de Francia, con lo que han tenido la posibilidad de practicar, aunque mínimamente, la lengua en un entorno endógeno.

6.2 Corpus CORAF: Resumen de los principales detalles técnicos

Nuestro corpus se compone de 30 entrevistas grabadas en formato *Waveform audio file format* o wav, y sus correspondientes transcripciones ortotipográficas en formato txt y xml. Ambos ficheros están, además, alineados sonido-texto con la ayuda del programa de transcripción *Transana*.

Las grabaciones audio se realizan con ayuda de una grabadora digital, que genera los ficheros directamente en wav, por lo que su almacenamiento y trasvase resulta muy fácil y bastante seguro. Posteriormente, dichos ficheros son tratados mínimamente con el programa *Adobe Audition*, sobre todo, con el fin de mejorar la calidad del sonido mediante la amplificación del mismo, y proceder a la eliminación de ruidos excesivos y sonidos de puesta en marcha y de apagado de la propia grabadora.

La transcripción y el posterior alineado se realizan una vez que los ficheros audio están revisados con la ayuda del programa *Transana*[99]. La transcripción, como veremos en capítulos posteriores de forma detallada, se realiza siguiendo las convenciones propias del Laboratorio de Lingüística Informática de la Universidad Autónoma de Madrid (LLI-UAM) para la realización de corpus orales, que tienen su origen en el formato CHAT, pero que han sido reformuladas y enriquecidas con el trabajo y la investigación llevada a cabo en sucesivos corpus, destacando, entre ellos, C-ORAL-ROM. Además, en el diseño de nuestro corpus se decide incorporar nuevas etiquetas, unas procedentes del corpus oral realizado también en el LLI-UAM con aprendientes de ELE y otras, exclusivas de nuestro corpus, que pretenden marcar los fenómenos propios de la lengua oral. En todo caso, la transcripción supone una ayuda importante para la comprensión del texto en caso de su utilización en comprensión oral, y es necesaria para todos los análisis, estudios y visualizaciones del corpus que queramos establecer.

Una vez transcritos y alineados, se realizan diferentes copias del fichero textual en distintos formatos, aunque se privilegia el uso del formato de texto plano o txt y de XML (*eXtended Markup Language*). La elección de ambos no es casual, ya que son aquellos formatos que permiten una mayor reutilización en diferentes herramientas y programas de explotación de corpus. Por su parte, hemos de señalar que la conversión a formato xml se realiza con ayuda de un programa específico en lenguaje de programación Perl concebido en el laboratorio LLI-UAM para su uso en distintos corpus orales.

Finalmente, una vez obtenidos los ficheros en el formato deseado, pudimos realizar los análisis previstos con ayuda de programas específicos como *UAMCorpus Tool*, *Wordsmiths Tools*, o sistemas de concordancias como *Contextes*, o *Microconcord*.

En nuestro caso, los ficheros sirvieron para realizar un análisis de errores de la interlengua del aprendiente hispanohablante de FLE, descrito más adelante en este estudio, basándonos en tres criterios: lingüístico, descriptivo y etiológico.

[99] Disponible en: http://www.transana.org/. El programa se comentará con más detalle en el capítulo 9 del presente estudio.

7. ¿Qué puede aportar el corpus CORAF?

Un corpus oral de aprendientes como el que aquí diseñamos puede, ante todo, convertirse en una fuente de datos no habitual para análisis de la interlengua. Con él, no sólo podemos estudiar los errores más frecuentes, sino realizar otros estudios sobre cuestiones relativas, por ejemplo, al discurso, al léxico más usado, al grado de utilización de fenómenos propios de la lengua oral, etcétera.

Hasta el momento, no conocemos ningún corpus oral o textual de aprendientes hispanohablantes de FLE difundido en el ámbito académico ni comercial de nuestro país, por lo que CORAF supone una novedad en este campo, abriendo nuevas posibilidades de exploración en la disciplina de FLE que no habían sido contempladas aún. Sí existen, sin embargo, estudios relativos a la interlengua del aprendiente hispanohablante de FLE, pero no se conocen los datos que sustentan dichos análisis.

Así las cosas, no debemos olvidar que nuestro corpus oral supone una primera aproximación, puesto que, lógicamente, requiere ser ampliada en lo sucesivo para lograr un corpus de mayor tamaño y número de palabras, que permita establecer generalizaciones válidas y consistentes para todo el conjunto de aprendientes al que concierne. Su tamaño reducido es, por tanto, una de sus primeras y más importantes limitaciones. Si bien, autores como Braun (2007) o Koester (O'Keeffe and McCarthy, 2010: 66), como hemos mencionado anteriormente, recuerdan que muchas veces, dependiendo del fin con el que se utilicen los corpus, es mejor un tamaño reducido, más manejable, que permita conocer a fondo su contenido estudiando todas las concordancias encontradas y no sólo las más frecuentes. Es el caso, por ejemplo, de usos directos del corpus en enseñanza con alumnos no iniciados en la materia (Cf. Flowerdew, 2002).

Nuestro corpus no servirá para estudios lexicográficos, que necesitan de una gran cantidad de datos para que aparezcan una cantidad determinada de expresiones o colocaciones. Sin embargo, como señala Koester (O'Keeffe and McCarthy, *op cit.*), ítems gramaticales como pronombres, preposiciones, verbos auxiliares y modales son muy frecuentes en el discurso y pueden observarse también en un corpus pequeño.

El hecho de compilar un corpus oral de aprendientes de lengua materna española permite también compararlo con otros corpus similares en tamaño y diseño realizados con otros aprendientes de distintas lenguas maternas, que sí han sido estudiados y difundidos en el ámbito de la adquisición de segundas lenguas. Por tanto, un estudio similar nos posibilitará realizar un análisis contrastivo de la interlengua (CIA), pudiendo desentrañar los procesos comunes a todos los aprendientes, y por consiguiente, universales, de aquellos específicamente ligados a la lengua materna (si los hubiere). Evidentemente, un estudio pormenorizado de la interlengua redundaría en un conocimiento amplio del aprendiente y en la concepción de un método de enseñanza específico y adaptado a las verdaderas necesidades y características de los aprendientes de nuestro entorno.

Por otra parte, cabe reiterar que nuestro corpus contiene muestras de interacciones orales con aprendientes de todos los niveles expuestos por el MCER, lo que supone una novedad con respecto a la mayoría de corpus disponibles, que se basan en un determinado nivel, generalmente, intermedio-avanzado. En este aspecto CORAF reúne a los corpus de aprendientes comerciales, que sí intentan establecer una visión total y amplia de la lengua meta. Por tanto, CORAF favorece una explicación de la interlengua en conjunto y en distintos niveles, posibilitando análisis de muy distinto tipo.

Asimismo, una de las críticas más frecuentes a los corpus de corte trasversal como CORAF es su ausencia de contexto (lo que les diferencia, generalmente, de los corpus tradicionalmente longitudinales). Apenas se tienen detalles del contexto donde se realiza la recogida de datos, ni incluso de los aprendientes que los configuran, lo que supone un problema para los teóricos e investigadores en adquisición de segundas lenguas, quienes los consideran datos esenciales.

Por consiguiente, para evitar esta posible limitación y por la importancia que conlleva para realizar análisis consistentes de la interlengua y de los errores de los aprendientes, en CORAF hemos decidido dar una mayor importancia al contexto de aprendizaje del aprendiente, a sus características y bagaje anterior y al propio contexto de interacción recogido en el corpus. Muchos de estos aspectos aparecen, por tanto, señalados en la cabecera de las transcripciones (con datos sobre el lugar y fecha de recogida de las muestras, información sobre el nivel del aprendiente, el tiempo aproximado de estudio de la L2, o el

conocimiento de otras lenguas, entre otros). Además, para complementar estos datos, se han incluido en las distintas entrevistas preguntas relativas al aprendizaje de la lengua meta (generalmente asociado al contexto guiado o institucional en el que se encuentran), varias cuestiones sobre su percepción de la L2 en sí y en relación con otras lenguas conocidas y/o estudiadas, así como reflexiones acerca de la lengua.

Todos estos datos nos servirán de ayuda en la elaboración de análisis rigurosos de la interlengua, donde se podrá dar cuenta de muchas variables que pueden influir en el proceso de adquisición, o en los errores y obstáculos detectados, y que no son habitualmente tenidas en cuenta en gran parte de los estudios por desconocimiento.

Finalmente, cabe reseñar que el formato escogido para la implementación de CORAF también resulta un aspecto positivo, puesto que su presentación en texto plano (txt) y xml favorece su reutilización y manejo en distintas aplicaciones y herramientas informáticas de explotación y de visualización de corpus. No obstante, hemos de señalar que el grado de eficacia demostrada por dichos programas no es del cien por cien, en parte, por el tipo de transcripción realizada y por la variedad de lengua que representa, que al contener aspectos diferentes de la lengua nativa, genera más inconsistencias con programas diseñados en su origen para otra variedad de lengua.

8. Conclusiones

A lo largo de este capítulo hemos intentado mostrar todos los aspectos que influyen y que han de tenerse en cuenta a la hora de diseñar un corpus oral de aprendientes, sobre todo, para lograr aspectos básicos como la representatividad, el equilibrio o una cantidad y adecuación de los datos al objetivo final de la investigación de la que forman parte. Hemos detallado cómo es nuestro corpus, qué es lo que contiene y qué podemos hacer con él.

Posteriormente, hemos descrito las dificultades previas que han motivado algunos cambios o influido directamente en el resultado final del corpus, destacando de todas ellas la escasa participación de EOI en nuestro proyecto, como contexto elegido para el desarrollo de las grabaciones, lo que repercute sin duda en un menor número de palabras

(y aprendientes), y por tanto, en una posibilidad menor de generalización y validación de hipótesis sobre la lengua.

Además, hemos desarrollado también una aproximación al corpus objeto de nuestro estudio, el corpus CORAF, señalando sus principales características, tanto de contenido como técnicas, finalidades y aportes al conjunto de la investigación, entre los que destaca su intención de suplir la carencia habitual de este tipo de recursos en el panorama académico y científico de nuestro país. No olvidamos así describir sus principales objetivos, aunque quizá el más inmediato sea el de proporcionar a los docentes e investigadores una nueva fuente de datos para la realización de análisis y estudios de la interlengua del aprendiente de FLE.

Entre sus posibles aplicaciones futuras destacamos la realización de distintos análisis contrastivos de la interlengua o investigaciones propias del campo de la adquisición de segundas lenguas. También es posible su uso directo, ya que su formato en txt y xml, aunque con algunas limitaciones, permite su reutilización en herramientas y aplicaciones de visualización y explotación de corpus. En concreto, este corpus servirá para la puesta en marcha de un análisis de errores frecuentes que pretende cubrir tres criterios distintos: lingüístico, descriptivo y etiológico.

Pese a su reducido tamaño, hoy por hoy, una de sus más importantes limitaciones, podemos afirmar que se trata de un corpus poco frecuente, ya que en el ámbito de FLE en España no hemos encontrado, por el momento, ningún corpus similar. Por lo tanto, CORAF supone una nueva contribución con respecto al tipo de corpus existentes en nuestro país, ofreciendo la posibilidad de llevar a cabo análisis desde enfoques metodológicos propios de la LC (Lingüística de Corpus). Análisis que, por otra parte, no son muy usuales todavía en las investigaciones para este campo de estudio.

Por tanto, con la intención de acercar los métodos y manuales generales existentes a los aprendientes de nuestro entorno inmediato, debemos de conocer sus especificidades, obstáculos y dificultades en la producción de la lengua meta. La forma aparentemente más sencilla de hacerlo es constituir un corpus de aprendientes, como se ha hecho para otras lenguas, con el fin de caracterizar la interlengua del aprendiente hispanohablante de francés como lengua extranjera y constituir mejoras pedagógicas en sucesivas aplicaciones relacionadas con las TICE y la enseñanza de lenguas asistida por ordenador (ELAO).

2. COMPILACIÓN E IMPLEMENTACIÓN DEL CORPUS CORAF

1. Introducción

Una vez concebido el diseño de nuestro corpus, hemos de ponerlo en práctica pasando a la fase de compilación e implementación. En este punto, todas las acciones que llevemos a cabo estarán determinadas por las características básicas con las que queremos dotar nuestro corpus y la metodología que hayamos decidido emplear para la recogida de los datos que lo conforman.

En la compilación de un corpus, se suceden distintos procesos, que iremos describiendo a lo largo de este capítulo y que se pueden resumir según el siguiente esquema:

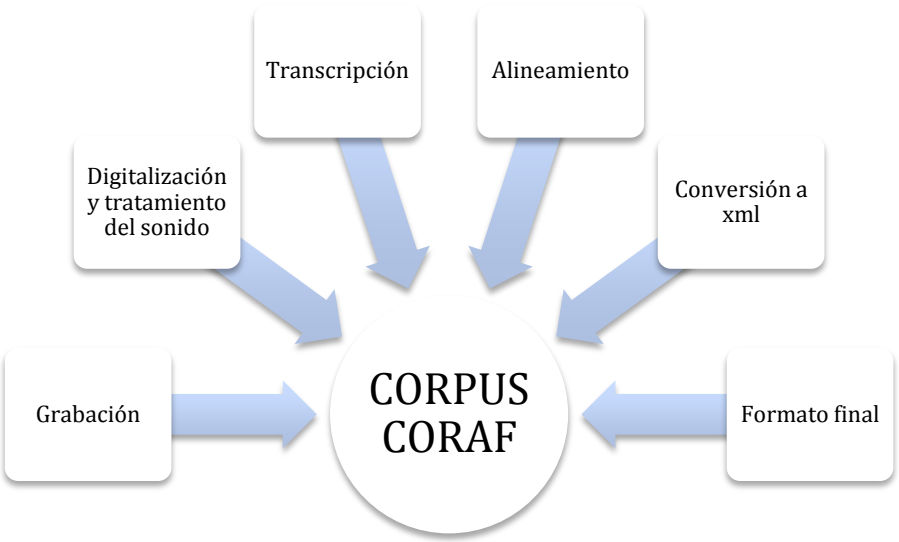

Gráfico 7: Procesos de implementación del corpus CORAF.

En primer lugar, haremos una descripción del proceso de grabación y de las dificultades surgidas durante el mismo. Posteriormente, nos referiremos a la digitalización de los datos obtenidos y al tratamiento de los mismos, y describiremos de forma detallada cómo se ha llevado a cabo la transcripción, deteniéndonos en las convenciones establecidas para su realización, así como en los fenómenos más reseñables que tuvieron que ser modificados para adaptarse al tipo de corpus al que nos enfrentamos. Nos ocuparemos también del alineamiento de texto y sonido y aludiremos a los procesos de conversión al formato deseado. Finalmente, mostraremos el aspecto final del corpus y comentaremos su contenido.

2. Grabaciones del corpus CORAF

La fase de grabación es quizá, junto con la concepción del diseño, una de las partes del proceso de implementación más importantes. De la cantidad y calidad de los datos obtenidos en ella depende, en gran parte, el valor de nuestro corpus y su potencial para que puedan derivarse de él distintas aplicaciones.

Ante todo, nuestra intención es conseguir unas grabaciones que se adapten a las características descritas en el diseño del corpus, de forma que resulten adecuadas para nuestro estudio de la interlengua, conteniendo datos representativos y suficientes. Por otro lado, queremos conseguir que las grabaciones posean una buena calidad, no sólo en cuanto a contenido se refiere, sino en aspectos técnicos como el sonido. Un sonido de calidad favorecerá su reutilización, no sólo en estudios basados en corpus, o en usos directos en entornos de aprendizaje o *data-driving learning*, sino en otras aplicaciones interesantes de la Lingüística Computacional basadas en el procesamiento del lenguaje natural.

2.1 Contextos de grabación y participantes

Como hemos indicado en el capítulo precedente, nuestra finalidad de caracterizar la interlengua de los hablantes de FLE nos lleva a escoger como contextos de grabación las Escuelas Oficiales de Idiomas (en adelante, EOI) y la Facultad de Letras de la UCLM.

Esta elección no es casual y responde no sólo al cumplimiento de los criterios básicos que debían de tener los aprendientes, sino a las ventajas que ofrecían dichas instituciones, a saber:

- Posibilidad de realizar grabaciones de muchos participantes pertenecientes a distintos niveles del MCER en el mismo momento y lugar, lo que redunda en una optimización del tiempo disponible;
- Mayor número de aprendientes dispuestos a colaborar al ser alumnos especialmente motivados para el estudio de una segunda lengua, y en contextos que garantizaran la calidad de la formación;
- Una mayoría de aprendientes adultos, por tanto con posibilidad de otorgarnos su autorización de forma inmediata, evitando así problemas legales posteriores o retrasos en la consecución del permiso.

Así las cosas, para llevar a cabo las grabaciones, nos pusimos en contacto con diez de las trece Escuelas Oficiales de Idiomas de nuestro radio de acción, explicándoles nuestro proyecto y pidiéndoles su colaboración en el mismo. De todas ellas, y tras varias comunicaciones, sólo tres accedieron a colaborar, dos de ellas en la provincia de Ciudad Real y una en Toledo[100].

Una vez obtenido el permiso para realizar las grabaciones en las tres EOI definitivas y en la Facultad de Letras, acordamos con los

[100] Hemos de señalar también que excepto una de ellas, en la que el jefe de departamento de francés tenía especial interés por el desarrollo de la competencia oral, las dos restantes se mostraron más receptivas al haber establecido previamente otras colaboraciones, para proyectos diferentes, con la Universidad de Castilla-La Mancha.

profesores y jefes de departamento respectivos las fechas en las que podíamos trabajar con los aprendientes, de forma que nuestro estudio no interfiriese demasiado con su actividad habitual.

Así, se acordó en la mayoría de los casos series de tres días consecutivos para las grabaciones de aquellos aprendientes de todos los niveles dispuestos a participar (por las tardes en el caso de las EOI, por las mañanas para la Universidad). Los profesores-tutores de los centros se encargaron de avisar previamente a sus alumnos del trabajo que queríamos realizar, de forma que a nuestra llegada, ya estaban formados los grupos de aprendientes voluntarios.

Las grabaciones se realizan, en la mayoría de los casos, en un espacio previsto de antemano para ello (en un despacho o en la sala de juntas), con la grabadora visible, y a solas con la entrevistadora/ investigadora[101]. Siguiendo la petición del profesorado colaborador, que no quiere que se interfiera en las clases de otros profesores, realizamos las grabaciones de los distintos niveles en la hora de clase que tienen asignada. Los alumnos van saliendo de su aula y acuden al lugar previsto para la grabación, volviendo a incorporarse a su clase una vez terminada la entrevista.

Las entrevistas tienen una duración diferente en función del nivel de los aprendientes. En un principio no establecimos ninguna duración mínima, aunque en la concepción del diseño creímos conveniente señalar como ideal una duración de entre doce y veinte minutos, ya que al querer realizar el máximo de grabaciones posibles en un espacio tan corto de tiempo, resulta imposible extender demasiado las entrevistas. En todo momento intentamos que el aprendiente respondiera a todas las preguntas, asegurando una duración mínima aproximada de doce minutos.

En los encuentros grabados, se fomenta una conversación continua, por lo que la grabadora no se para en ningún momento hasta que no concluye la entrevista. En ningún caso ha sido necesario una pausa en la conversación, pudiendo los participantes completarlas con éxito.

[101] En el caso concreto de una de las EOI, las grabaciones son hechas cambiando de entrevistadora. Habitualmente la investigadora realiza también las entrevistas, pero en este caso, por política del centro, tres estudiantes en prácticas del Máster de Enseñanza Secundaria son las que se encargan de realizarla (una hablante nativa, otra bilingüe y una tercera no nativa), estando la investigadora presente.

Como hemos indicado anteriormente, no sólo se les pregunta por aquellas cuestiones ya previstas en el esquema de la grabación, y que posibilitan una coherencia intertextual del corpus, sino que se establece una verdadera conversación añadiendo nuevas preguntas o modificándolas en función de las respuestas del aprendiente. Todo ello conforma un conjunto de muestras que llegan a alcanzar, en el caso de los aprendientes del nivel competente (C1 y C2) hasta los veinte minutos. Así, la duración de las muestras va aumentando en función del nivel, siendo la media de todas ellas de 15 minutos y 09 segundos.

Al principio todos los participantes se muestran algo nerviosos ante la situación, restando espontaneidad a sus respuestas. Sin embargo, pasados unos minutos, y coincidiendo normalmente con el final de la primera fase de la grabación donde realizan una pequeña presentación de su vida y entorno, la mayoría de los hablantes se olvida de la situación y de sus características (investigador desconocido que interactúa con ellos, grabadora visible que registra su conversación, novedad de la tarea que han de realizar) y hablan normalmente, como en cualquier otra conversación habitual en un entorno nativo. Evidentemente, este hecho es más apreciable en los niveles más avanzados de conocimiento de lengua, pues la propia inseguridad de los hablantes de A1 o A2, que todavía, habitualmente, no han tenido la oportunidad de expresarse extensamente en la lengua meta, les lleva a ceñirse a la conversación propuesta por el investigador, a no salirse del "guión" establecido y a esperar sus preguntas, respondiéndolas escuetamente y con una menor creatividad, con la intención de demostrar que conocen o que han entendido lo que se les pregunta, pero sin aportar nada nuevo que les pueda inducir a errores.

Los datos relativos a la duración total y media de las entrevistas realizadas, así como el número de participantes total de la fase de grabación de nuestro proyecto, se pueden observar en la siguiente tabla:

NIVEL MCER	Nº ENTREVISTAS	DURACIÓN TOTAL	DURACIÓN MEDIA	LOCUTOR (M/H)
A1	13	2h 32m 56s	11m 46s	10M / 3H
A2	13	3h 00m 59s	13m 55s	6M / 7H
B1	6	1h 34m 22s	15m 44s	5M /1H
B2	7	1h 54m 13s	16m 19s	5M / 2H
C1	11	3h 06m 30s	16m 57s	9M / 2H
C2	12	3h 14m 16s	16m 11s	7M / 5H
TOTAL	62	15h 23 16s	15m 09s	42M / 20H

Tabla 8: Resumen de los datos principales del contenido de las grabaciones del corpus CORAF.

Los participantes colaboran generalmente con mucho empeño con nuestro proyecto, logrando grabar 62 aprendientes, un total de 42 mujeres y 20 hombres, con distintos perfiles sociodemográficos, que se reflejan de la forma siguiente:

- Edad: Los participantes se agrupan en 17 personas de entre 18 a 25 años (rango A), 20 de entre 25 y 40 años (rango B), 22 de entre 40 y 60 años (rango C) y 3 menores de 18 años.

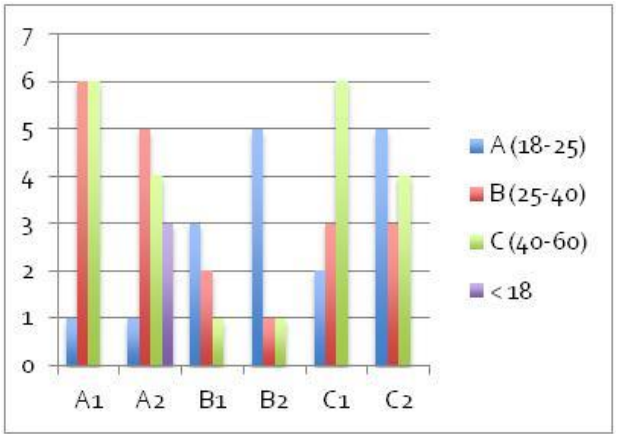

Gráfico 8: Reparto de rangos de edades en CORAF por niveles del MCER.

- Nivel de estudios: un participante cuenta con estudios básicos (nivel 1), 11 con estudios secundarios (nivel 2) y 50 con estudios universitarios (nivel 3).

Como venimos comentando, las grabaciones resultantes no están equilibradas en todos los niveles del MCER, al no haber encontrado a alumnos suficientes para el nivel B1 y B2 (quedando sólo con 6 y 7 participantes, respectivamente). Además, a este desequilibrio hemos de sumarle otro similar en el nivel A2, donde hay otros tres participantes que son menores de edad, reduciéndose de 13 a 10 el número de grabaciones. Aunque en el momento de difundir este estudio, son ya adultos, preferimos no utilizar dichas muestras al no contar con la autorización paterna para su participación (aunque sí el permiso firmado de los tres aprendientes).

Por consiguiente, modificamos nuestro diseño inicial y optamos por reducir el número de muestras que serán transcritas a cinco, debido sobre todo, al límite que nos impone el nivel B1 con tan sólo seis muestras.

Hemos de reconocer que la cifra de participantes es demasiado reducida, y que debido a ello, las hipótesis que podamos generar sobre la interlengua no serán extrapoladas al conjunto de los aprendientes de FLE. Ante la imposibilidad de llevar a cabo nuevas grabaciones, y debido

a que la duración de las entrevistas era lo suficientemente importante para conseguir datos que avalen nuestro posterior análisis de errores (como veremos en el siguiente capítulo), hemos optado por continuar nuestro estudio, siendo conscientes de que será, sin duda, una primera aproximación o descripción de los comportamientos habituales de los hispanohablantes que aprenden FLE en contexto académico. De modo que el presente estudio requiere ser completado en un futuro próximo, y con nuevas grabaciones, con un número de aprendientes participantes igual o superior a diez para cada uno de los niveles.

2.2 Aspectos técnicos generales

Las grabaciones se realizan con una grabadora de audio digital EDIROL R-09HR, de la marca Roland. Dicha grabadora permite llevar a cabo grabaciones con mucha calidad en formatos wave y mp3, de 16 ó 24 bits a través de un potente micrófono estéreo.

De sus características técnicas principales, destacamos dos: la posibilidad de ajustar fácilmente la calidad del sonido que deseamos captar durante la grabación y su fácil exportación a otros dispositivos.

Así, en cuanto a la calidad del sonido, se puede elegir la frecuencia de muestreo, el volumen de entrada de la grabación y el tipo de formato de salida del archivo. En nuestro caso, optamos por grabar con una frecuencia de 48 Khz. a 16 bits, que supone una calidad media y que permite poder realizar posteriormente un tratamiento del sonido, si fuera necesario, sin merma de calidad. Como formato del archivo elegimos .wav, que almacena todo el sonido sin compresión (a diferencia del mp3), lo que proporciona una mayor calidad.

En cuanto a su facilidad para almacenamiento y exportación de archivos, debemos de señalar que el hecho de que se puedan almacenar los datos en una tarjeta de memoria de tipo SD (*Secure Digital*), simplifica mucho la tarea. De esta forma, para nuestras grabaciones, utilizamos dos tarjetas de sonido de alta calidad, SD HC, de 4 GB, donde se recogen las muestras para después exportarlas a otros dispositivos de almacenamiento.

La grabación se realiza en un espacio aislado (aunque esto no evita la aparición de muchos ruidos debido a las características de los

edificios, como explicaremos más adelante), donde el participante se encuentra, como ya dijimos, con el investigador a solas. La grabadora se sitúa cerca de los interlocutores, en un soporte especial a modo de trípode que la aísla de las vibraciones producidas por movimientos de los participantes que se puedan transmitir a través de la mesa en la que se encuentra, perjudicando la grabación. Debido a su reducido tamaño, similar al de un *smartphone*, no supone un elemento molesto, y permite, que pasados unos minutos, el participante no sea consciente de su existencia.

Los archivos de audio con las entrevistas de los aprendientes, como veremos más adelante, se exportan y se almacenan como formato wav para garantizar su mayor calidad e interoperabilidad con programas del tratamiento del sonido. La versión final de los archivos de CORAF será ofrecida en .wav, y eventualmente en .mp3 si es necesaria su incorporación a sistemas o plataformas informáticas, ya que es un formato de menor tamaño y más fácilmente asumible por las distintas aplicaciones y herramientas.

2.3 Grabaciones: dificultades y limitaciones

Las principales dificultades que hemos encontrado en la puesta en marcha de la fase de grabación no son numerosas, sin embargo tienen una repercusión importante en el contenido del corpus, y pueden resumirse de la siguiente forma:

- Ausencia de espacios aislados para la realización de las entrevistas: Pese a la buena voluntad de los centros educativos visitados, encontramos problemas para poder desarrollar nuestras grabaciones con una total ausencia de ruidos. Primero, porque las características de los edificios, algunos muy antiguos, no garantiza un correcto aislamiento de las aulas o salas. Todo ello produce una fuerte presencia de ruidos provenientes del exterior en nuestras grabaciones (ruido de una obra cercana, audio de las actividades de comprensión oral de otras clases, alarma de final de clase del centro, pájaros que vuelan cerca de las ventanas, etcétera). Segundo, porque existen centros donde no se disponía de un aula concreta que pudiera ser ocupada, lo que suponía ir cambiando de ubicación según la hora, e incluso, no disponer de ningún aula y

tener que realizar las grabaciones en un pasillo aislado. Todo ello repercute en la calidad de las grabaciones, y unido al problema de desequilibro en el número de grabaciones para algunos niveles, en la decisión de centrar nuestro corpus en sólo 30 entrevistas. En todos los casos, los ruidos y los eventuales problemas de sonido que puedan aparecer son comentados en la cabecera de las transcripciones y en las partes de la misma donde interfieren con el audio de la interacción, de forma que el usuario puede conocer en todo momento la calidad de las grabaciones que va a escuchar.

- Limitación de tiempo para realizar las grabaciones: Nuestra intención es interferir lo menos posible en el trabajo realizado en las EOI y la facultad, por lo que llevamos a cabo las grabaciones de forma continua en los días que nos asignan para ello[102], lo que supone realizar el mayor número de grabaciones en el menor tiempo posible. Por tanto, la duración de algunas muestras, que podía ser más extensa por la disposición del aprendiente a la interacción, tiende a verse reducida al tener alumnos esperando para realizar las entrevistas, o pendientes de sus clases. En otros casos, exige también una programación de entrevistas cada quince minutos para que todos puedan participar.

- Dificultad para encontrar hablantes de nivel B1 y B2: El hecho de no obtener suficientes aprendientes del nivel independiente (nivel B) motiva, como ya hemos mencionado, la limitación del corpus oral transcrito a sólo cinco entrevistas por nivel. Las causas de esta ausencia son principalmente tres:
  - La ausencia de cursos programados al carecer de alumnos suficientes en algunas de las EOI, y de alumnos en las clases presenciales de Filología Francesa;
  - La escasez de alumnos voluntarios para ser entrevistados, bien porque no quieren perder tiempo de su instrucción, bien

[102] Hay que señalar que algunos profesores no quieren que sus alumnos pierdan clase, con lo que las grabaciones se realizan durante esa hora y van entrando y saliendo del aula mientras se desarrolla, lo que condiciona significativamente la duración de la grabación.

porque no se encuentran cómodos ante la tarea que han de desarrollar, desconfiando de su nivel y competencia;

- La negativa por parte de algunos de los profesores-tutores encargados de dichos niveles a participar en nuestro proyecto por la presión del tiempo.

- Ausencia de centros dispuestos a colaborar con nuestro proyecto, de forma que tuviésemos un mayor número de aprendientes con los que poder realizar nuestros estudios, aumentando la representatividad del corpus CORAF.

Si bien no todas estas dificultades han podido ser superadas en las sucesivas grabaciones del corpus CORAF, al menos servirán para ser tenidas en cuenta para futuras ampliaciones y revisiones del mismo, de forma que podamos aumentar progresivamente la validez y calidad del corpus.

3. Digitalización y tratamiento del sonido

En la fase de digitalización nos ocupamos de extraer las muestras recogidas en la grabadora para almacenarlas en otro dispositivo, generalmente un ordenador, con el fin de poder tratarlas y trabajar en ellas con distintas aplicaciones y herramientas informáticas.

La digitalización fue para nosotros un proceso muy sencillo, puesto que las nuevas grabadoras digitales evitan el arduo trabajo que imponían las antiguas máquinas, consistente en pasar los datos de las cintas a través de un *software* especial al ordenador. Con las grabadoras digitales basta con pasar a través de un cable USB los datos al ordenador o soporte elegido, archivándolos y dándoles un nombre en función de nuestras necesidades.

En CORAF, agrupamos las entrevistas de cada nivel del MCER en carpetas diferenciadas y les atribuimos un nombre que especifica el nivel al que pertenecen (A1/A2/B1/B2…), el género del participante (Woman para mujer-W- y Man para hombre-M-) y un número que le

distingue del resto, en orden correlativo. Por ejemplo: A1W02, B2M01 ó C2W03.

Una vez almacenadas en formato .wav, realizamos en ellas un leve tratamiento para mejorar su calidad. Aunque nuestra grabadora proporciona unas muestras de excelente calidad, nuestras grabaciones no dejan de ser situaciones comunicativas reales, por lo que existen elementos que no podemos controlar o solucionar como si estuviésemos haciendo grabaciones controladas en un estudio. Así, es habitual que de vez en cuando aparezcan ruidos inesperados provenientes del entorno o que el participante habla con un tono muy bajo. En todos estos casos es necesario realizar un tratamiento del sonido con un programa de edición.

Para mejorar la calidad del sonido y darle forma a los ficheros de audio utilizamos el programa *Adobe Audition*. El tratamiento que se hace de todos los ficheros es mínimo, pero existen varios procesos que se realizan en todos el conjunto:

- Cambio de la frecuencia de muestreo de 48 Khz. a 22 Khz., con el fin de que los archivos wav sean más fácilmente recuperables por los programas informáticos que nos ayudan en la transcripción.
- Adaptación de la muestra, suprimiendo los segundos relativos a la puesta en marcha de la grabadora y al de pausa final de la grabación, para evitar el ruido que provoca el presionar el botón de encendido y parada.
- Si el archivo no tiene la suficiente calidad, sobre todo, si tiene un volumen muy bajo, llevamos a cabo la amplificación del fichero para mejorarlo.
- Si el archivo posee algún ruido fuerte, susceptible de provocar molestias en los futuros oyentes, tratamos de suprimirlo reduciendo la amplificación de esa zona de onda concreta. En nuestro proyecto se ha realizado esta acción con carcajadas muy sonoras, ruido de una puerta que se cierra de golpe y la caída de un libro al suelo.

Una vez realizado el tratamiento del sonido, los archivos de audio en formato .wav están listos para iniciar la siguiente fase, que es, sin

duda, una de las más importantes de toda la compilación de un corpus: la transcripción.

4. Transcripción

Los corpus orales tienen en su proceso de transcripción uno de sus mayores desafíos. Es evidente que representar la lengua oral no es un trabajo fácil, ya que la riqueza que suele tener el discurso oral supone una dificultad y un gran esfuerzo para el investigador si decide representarla tal y como es.

Una de las razones de su dificultad reside en su carácter multimodal, agrupándose en la interacción oral elementos no sólo textuales, sino también prosódicos, gestuales y otros derivados del contexto de la situación comunicativa, que inciden en la construcción de la misma y de su significado.

Un corpus oral necesita ser transcrito para ser utilizado en cualquiera de sus fines, ya sean científicos o de investigación como educativos. En realidad, y debido al tipo de herramientas que se manejan habitualmente en la LC, lo que se explota y visualiza es realmente el texto (aunque pueda ir acompañado del audio), ya que, como hemos comentado anteriormente, las aplicaciones que procesan la lengua oral no están tan desarrolladas como las existentes para la lengua escrita.

Además, es la fase de creación de un corpus oral más costosa ya que no existen herramientas que puedan sustituir el trabajo del investigador, siendo por tanto, una tarea completamente manual, que requiere de una alta concentración y de una gran cantidad de tiempo, ya que, para media hora de transcripción podemos llegar a necesitar unas diez o doce horas de trabajo. Es, sin duda, una de las causas que motivan la escasa compilación de corpus orales, mucho menos frecuente que para los corpus escritos.

Así, con el fin de explotar fácilmente nuestro corpus en un futuro, necesitamos transcribirlo. En este punto es necesario reflexionar acerca del nivel de detalle necesario para esta transcripción, lo que derivará en criterios y pautas específicas para su desarrollo. También es necesario seguir ciertos criterios estandarizados, de forma que el corpus pueda ser reutilizado por la comunidad científica. Sin embargo, todavía no se ha

alcanzado ningún acuerdo acerca de un estándar de transcripción (pese a que muchos estudios utilizan las pautas de CHAT para sus corpus), lo que nos obliga a utilizar aquellos estándares más frecuentes de nuestro entorno[103].

En realidad, el nivel de detalle de la transcripción dependerá del objeto de nuestra investigación. Así, en nuestro caso deberemos de atender, además de a una fiel transcripción ortográfica, a dos aspectos concretos:

- Transcripción fiel de los errores presentes en la interacción, evitando su corrección o modificación y reflejando, de manera aproximada, si no existe la palabra o expresión, lo que quieren decir, así como no señalando las *liaisons* obligatorias u otras contracciones de formas lingüísticas o *élisions (*je ai'* en lugar de *'j'ai'* ó *'*à la école'* en vez de *'à l'école'*).
- Marcado de los fenómenos propios de la lengua oral, es decir, considerados como propios de de dicho registro, lo que nos permitirá ver el grado de apropiación de la lengua meta y realizar distintos estudios sobre el desarrollo de esta en aprendientes.

Para reflejar el conjunto de la interacción oral, optamos también por reflejar aspectos prosódicos, los signos extralingüísticos, así como otros no lingüísticos propios de esta como reformulaciones, pausas, solapamientos… que nos ayudan a comprender las características propias de la interacción y aportan una mayor riqueza al corpus, que podrá ser analizado desde otros puntos de vista distintos al puramente descriptivo o lingüístico, posibilitando así análisis pragmáticos, psicolingüísticos y de análisis del discurso.

Todas las transcripciones se realizan con ayuda de un programa informático, **Transana**[104], de análisis de archivos de datos digitales (audio y/o video) creado por Chris Fassnacht, pero mantenido en la

[103] En los últimos corpus encontramos, al menos para algunos corpus de aprendientes significativos como FRIDA, el uso de estándares de TEI (*Text Encoding Initiative*) para transcripción del habla (http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TS.html). Algo que deberíamos quizá plantearnos utilizar para una mejora posterior de CORAF, ya que sería más fácil su difusión y reutilización.
[104] http://www.transana.org/

actualidad por David K. Woods en el *Center for Education Research* de la Universidad de Wisconsin-Madison.

Transana es un programa muy sencillo, que permite analizar y manejar los datos de forma diversa, pero sobre todo, permite realizar la transcripción del sonido de forma fácil al mantener en la misma ventana todos los archivos concernientes (audio/video y texto), de forma que el transcriptor no tiene que salir del programa para escuchar el sonido, algo que antes ocurría con la mayoría de herramientas similares.

Otro de sus aspectos positivos es que es un programa muy ergonómico, ya que posee distintos atajos a través de comandos del teclado que permiten al transcriptor manejar el audio sin necesidad de coger el ratón, de forma muy rápida y sencilla. Así es posible retroceder, parar, adelantar o reanudar la audición, entre otros, con una serie de comandos a partir de la tecla Control (Ctrl) y la letra concerniente.

Aunque hasta hace poco tiempo era un programa gratuito, en la actualidad es necesario poseer una licencia para su instalación. Esta licencia no es muy costosa, por lo que es fácil de asumir y permite, sobre todo, ahondar en otra de sus características principales: sus versiones multiplataforma, de forma que podemos utilizar el programa en distintos sistemas operativos como Windows y Mac OSX

Finalmente, otro de los motivos por los que elegimos *Transana* es que permite realizar a la vez que se transcribe el alineamiento del texto con el sonido, introduciendo una marca de tiempo en los momentos en que el transcriptor lo desea con el comando control + T, de forma muy sencilla y rápida. Además, una vez realizado el alineamiento nos permite visualizar los segmentos de tiempo ya alineados señalándonos con un tono azul las partes de texto que se corresponden con el audio que estamos reproduciendo.

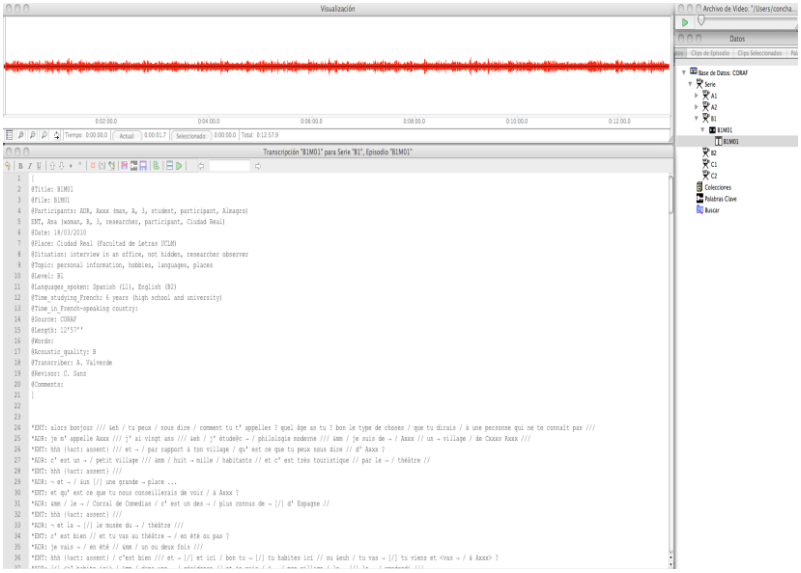

Ilustración 1: Captura de pantalla del trabajo de transcripción con el programa Transana (versión para Mac OSX)

4.1 Convenciones y pautas de transcripción

Para que la transcripción se realice de la forma más detallada posible, ha de contar con una serie de convenciones y normas que garanticen una coherencia en el desarrollo del proceso, de forma que contemos con el mismo tipo de transcripción para todos los archivos del corpus.

Como hemos comentado anteriormente, existen varios intentos de estandarización de las pautas de transcripción, como las provenientes del sistema CHAT[105] o la iniciativa TEI[106] (*Text Encoded Initiative*), pero en la realidad, y aunque suelen utilizarse para diseñar los aspectos básicos de las normas de transcripción, los investigadores suelen acomodar o cambiar sus pautas en función del corpus que realicen o según las

105 http://childes.psy.cmu.edu/
106 http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TS.html

necesidades de su investigación. Más aún cuando se trata de un corpus oral o de un corpus de aprendientes, que conlleva fenómenos especiales o particulares que no han sido tratados en estas convenciones estándar.

En nuestro caso, realizamos una transcripción ortográfica lineal cuyas convenciones han sido ya utilizadas y progresivamente mejoradas en otros corpus de lengua oral anteriores, entre los que destacan:

- El formato utilizado en el corpus C-ORAL-ROM (Cresti y Moneglia, 2005), que se apoya en el sistema CHAT;
- Los corpus CHIEDE y Corpus de aprendientes de ELE realizados en el Laboratorio de Lingüística Informática de la UAM;
- Las pautas de marcado utilizadas en el seno de la base de datos FLLOC, corpus de interlengua francesa (Rule, Marsden, Myles and Mitchell, 2003).

Sin embargo, para subrayar las especificidades de nuestro corpus, añadiremos o cambiaremos algunas normas de transcripción, que serán detalladas más adelante en este capítulo.

Finalmente, es importante señalar que las transcripciones se componen de dos partes diferenciadas:

- Una cabecera que contiene todos los datos acerca de la grabación y de los participantes.
- El texto que podemos escuchar en el archivo de audio de la grabación transcrito según las normas y pautas especificadas para nuestro corpus.

4.1.1 La cabecera de la transcripción

Al comienzo de cada uno de los archivos de texto de CORAF encontramos una cabecera que contiene la información esencial sobre el texto al que acompaña. Aparece entre corchetes, para separarla así del resto del texto, y está redactada en inglés siguiendo las convenciones internacionales, y sobre todo, lo expuesto en el proyecto C-ORAL-ROM (Cresti y Moneglia, 2005).

Esta cabecera incluye información general sobre la situación comunicativa, es decir, sobre el contexto en el que se desarrolla, y sobre los participantes, atendiendo sobre todo, al aprendiente, del que describe aspectos demográficos y de su experiencia lingüística.

Al tratarse de un corpus de aprendientes creímos conveniente destacar en la cabecera aspectos esenciales relativos a la interlengua, como el grado de conocimiento de la lengua meta (expresado a través del nivel del MCER en el que se sitúan), años de formación, conocimiento de otras lenguas o estancias prolongadas en países donde se habla la lengua meta[107]. Todos estos datos nos servirán de ayuda para comprender parte del proceso de adquisición y en el momento del análisis de la interlengua.

Los datos que contiene esta cabecera se pueden detallar de la siguiente forma:

@Title: Título que puede describir la grabación. En nuestro caso, escogimos utilizar el nombre del archivo de audio al que se refiere (por ejemplo, B2W01).

@File: Nombre del archivo con el que diferenciamos a cada entrevista, que se compone de la combinación entre el nivel en términos del MCER en el que se sitúa el aprendiente, el código de género del participante (W para mujer, M para hombre) y un número de dos cifras correspondiente al número de entrevista, que lo diferencia del resto de grabaciones del mismo nivel. Por ejemplo: B2M01, donde encontramos un participante de nivel B2 (avanzado), de sexo masculino y cuya entrevista tiene el número 01.

@Participants: Detallamos aquí toda la información relativa a los hablantes que participan en la grabación, y se incluye:

- Código de tres letras mayúsculas que identifica al participante y que servirá para marcar los distintos turnos del hablante en la transcripción. En nuestro caso, para el aprendiente, utilizamos las tres primeras letras de su nombre (por ejemplo, ALR). Si existen dos participantes en los que coinciden las letras, utilizamos las iniciales de los apellidos. Para el

[107] Seguimos así la metodología utilizada en distintos corpus de aprendientes europeos y en el corpus de aprendientes de ELE actualmente compilado en el LLI-UAM.

entrevistador, si se trata del investigador, utilizamos ENT, si no, guardamos las mismas normas que para el aprendiente.

- Nombre del hablante y entre paréntesis, sus principales características sociodemográficas. (El nombre se anonimiza dejando únicamente su inicial, y si desconocemos alguna de sus características, se señala con una 'x'). Se compone de:
  - Sexo: man / woman
  - Edad: expresada a través de los siguientes códigos:
    - A: 18-25 años
    - B: 25-40 años
    - C: 40-60 años
    - D: más de 60 años
  - Nivel de estudios: 1, estudios primarios; 2, estudios secundarios y Bachillerato; 3, estudios universitarios (graduados, licenciados o doctores)
  - Profesión: ocupación o trabajo que desarrolla habitualmente
  - Papel que desarrollan en la grabación (participante, observador…)
  - Origen geográfico

Podemos observar cómo se expresan estas características en el siguiente ejemplo: ALR, Axxx (man, B, 3, teacher, participant, Toledo)

- @Date: Fecha en la que tiene lugar la grabación.
- @Place: Lugar en el que se realiza la grabación.
- @Situation: Descripción de las características de la situación comunicativa, aludiendo al tipo de interacción (en nuestro caso, una entrevista), en qué lugar se desarrolla y con qué condiciones especiales (generalmente, si la grabadora está visible o no, o si el investigador se encuentra presente en la interacción).
- @Topic: Describe los temas en torno a los que gira nuestra grabación.

- @Level: Indica en qué nivel en términos del MCER se encuentra escolarizado el aprendiente.

- @Languages_spoken: Expresa las lenguas que conoce el aprendiente y su grado de conocimiento en términos del MCER, si se tiene constancia. Se incluye también aquí su lengua materna.

- @Time_studying_French: Indica el tiempo que el aprendiente lleva estudiando francés, en meses o años, y datos importantes relativos a su instrucción, como si se realizó en el instituto o sólo en la EOI, o bien en otro tipo de entidades.

- @Time_in_French-speaking country: Alude al tiempo de estancia continuada en países francófonos del aprendiente, expresado en meses/años y detallando el lugar o país visitado.

- @Source: nombre del propietario que ostenta los derechos de reproducción.

- @Length: Duración de la grabación expresada en minutos (') y segundos (").

- @Words: Número de palabras totales de la transcripción y número de estas producidas exclusivamente por el aprendiente.

- @Acoustic_quality: Calidad acústica de la grabación, expresada con el código A (grabación digital), B (grabación analógica con ruido de fondo) y C (grabación analógica de mala calidad).

- @Transcriber: Identificación de la persona que realizó la transcripción.

- @Revisor: Persona que se ocupa de revisar la transcripción.

- @Comments: Comentarios generales sobre la transcripción que sean significativos y puedan interesar para futuras aplicaciones o para entender la situación comunicativa. Suelen expresarse comentarios acerca de circunstancias particulares que pueden afectar la grabación, como ruidos, cambios en la situación comunicativa (aparición de nuevos participantes durante la grabación o cambio de entrevistador) o comentarios sobre los aprendientes, en particular, sobre su expresión (problemas en la

expresión de algunos fonemas, o acentos propios de su variante dialectal, etcétera).

4.1.2 Texto de la transcripción

CORAF recoge transcripciones ortográficas de interlengua sin tener en cuenta las convenciones tipográficas habituales para la lengua escrita. Sin embargo, demuestra una gran riqueza al incidir en otros aspectos específicos de la interacción, codificados a través de distintas marcas, como los siguientes:

- Aspectos prosódicos, indicando las pausas entonativas y las marcas de continuación de turno, enunciado suspendido o final del enunciado.

- Fenómenos propios de la oralidad espontánea como repeticiones, reformulaciones, apoyos vocálicos, alargamientos, autointerrupciones, solapamientos, signos paralingüísticos o fenómenos característicos y exclusivos de la variedad oral de la lengua meta (como cambios en el orden canónico, introducción de palabras de registro conversacional, etcétera).

- Fenómenos relacionados con el análisis de la interlengua como errores de conjugación, problemas significativos de pronunciación, formas o palabras inventadas, uso de palabras de otros idiomas o repetición/imitación de las fórmulas o expresiones utilizadas previamente por el entrevistador.

- Comentarios específicos relacionados con distintas expresiones, errores, problemas o aspectos de la interacción.

Por otra parte, hemos de recordar que respetamos en todo momento el discurso del aprendiente, que se transcribe tal y como se escucha en la grabación, lo que supone incluir los errores que puedan cometer. Así, también intentamos reproducir de la forma más realista posible todas las formas que no existen o

que inventa, realizando, en ocasiones, una transcripción fonética de lo expresado.

Finalmente, en cuanto a las convenciones de escritura:

- eliminamos los guiones entre palabras, excepto para las palabras compuestas,
- mantenemos un espacio siempre antes y después de cada palabra, lo que supone separar grupos de palabras que se contraen (como pronombre personal sujeto de primera persona y verbo conjugado que empieza por vocal)

4.1.2.1 Turnos

Para marcar los distintos turnos que componen la interacción oral, utilizamos el código de tres letras que identifica a cada participante en la cabecera, precedido de un asterisco y seguido de dos puntos:

*ENT: hhh {%act: assent} / tu travailles // et → qu' est ce que tu fais comme → / travail ?

*ALR: &eh / je suis professeur / &eh / de [/] dans l' école ///

[A1M01]

4.1.2.2 Etiquetas de información prosódica

4.1.2.2.1 Pausas entonativas

El participante, a lo largo de su discurso, puede realizar pequeñas pausas tonales sin que ello comporte el final de su enunciado. Para marcar las distintas unidades tonales utilizamos dos símbolos: / (barra simple) y // (barra doble).

La barra simple (/) se utiliza en pequeñas pausas tonales, donde generalmente el hablante realiza una leve inflexión tonal, que nos induce a pensar que todavía no ha terminado su enunciado.

La barra doble (//) señala una pausa tonal más marcada, pudiendo incluir a veces una entonación descendente, y donde encontramos unidades tonales con un mayor significado, pero que no componen, de ninguna manera, el fin del enunciado, ya que el hablante prosigue su discurso. La diferenciación entre pausa entonativa de barra doble y de final de enunciado no es muy clara en ocasiones, y queda casi siempre, a merced del criterio del transcriptor y/o revisor. Esta marca suele formar parte de enumeraciones, aclaraciones o incisos, etcétera.

Podemos ilustrar ambas marcas con un ejemplo:

*ANM: et j' ai → resté un peu // et après → / j' ai → / étudié /// &eh / &mm / &eh / j' ai écouté → / la [/] de la musique // et j' ai → parlé / à [/] au téléphone // et je suis → / &eh / &aran [/] dormi → à → [/] &eh [/] onze heures // de la nuit ///

*ARI: hhh {%act: assent} /// tu m' as dit que tu → / écoutes de la musique // quel genre de musique / tu aimes ?

[A2W01]

4.1.2.2.2 Marcas de continuación de turno y final de enunciado

El final de enunciado puede marcarse a través de los símbolos ?, !, ///, … , +, indicando así, respectivamente, un enunciado interrogativo, exclamativo, asertivo, enunciado inacabado o interrumpido por otro hablante.

La triple barra (///) señala el final de un enunciado asertivo, generalmente marcado por una entonación descendente:

*SAN: &eh / parce que → / &esa → [/] &eh / il est → / &ah [/] intéressant pour moi ///

[A1M02]

Los signos de exclamación (!) e interrogación (?) señalan el final de un enunciado exclamativo e interrogativo. Ambos signos se separan de la palabra anterior por un espacio, contrariamente a lo que establecemos en la lengua escrita, de forma que podamos diferenciarlos de las interjecciones (*ah!, bof!*) o de marcadores conversacionales de confirmación (*non?, hein?*), que se marcan con dichos signos, pero sin espacios.

*ENT: et Sxxx / vous → [/] vous → êtes née ici ?

*SUB: non je suis née à → / Paris ///

[A1W03]

En este punto, debemos indicar que hemos realizado algunas modificaciones en lo que a enunciados interrogativos se refiere. En ocasiones encontramos varias preguntas seguidas, que se complementan entre sí, o en las que se muestran incisos donde se realiza algún tipo de matización, explicación o especificación. Es evidente que al ser enunciados interrogativos deberían de ser enunciados distintos y señalarlos como tal, con la marca de final de enunciado, pero en realidad, y teniendo en mente el alineamiento posterior, hemos preferido no separar estos enunciados, puesto que corresponden al mismo, y en ocasiones, son necesarios para comprenderlo en su totalidad. En estos casos hemos optado por señalar estos enunciados como barra doble (//), o pausa entonativa, de forma que continúa el enunciado hasta finalizar con la última pregunta del mismo, donde ya se señala con el signo de interrogación.

*ENT: d'accord /// et → / du français / qu' est ce que c' est le plus difficile // il y a quelque chose qui → soit difficile pour toi // ou il y a [/] il n' y a rien ?

[A2M02]

La marca de enunciado inacabado (…) generalmente señala algo que bien el hablante no puede acabar porque tiene dudas, o porque cree expresar algo que el oyente puede deducir del contexto, acompañándose entonces de una entonación suspendida:

*STE: [<] <hhh {%act: laugh}> /// donc / &eh / si tu → [/] si tu pouvais me décrire un lieu en France que → [/] que tu as visité / que tu as aimé / <ou → > ...

*ANT: [<] <&ah> / j' ai visité → / Paris mais → / la première fois que je suis allé en France // je → / suis allé à → [/] à &Ager [/] <&Agé [///] &Angé xxx> ?

[C1M02]

La marca de interrupción por otro hablante (+) se utiliza para indicar un enunciado que no ha sido terminado porque el oyente interrumpe al hablante. Es muy habitual en conversaciones y forma parte de los rasgos de espontaneidad de este tipo de interacción:

*PSU: ah! oui / voilà /// mais → / hhh {%act: click} / finalement → / je → [/] j' ai décidé de → [/] de suivre / dans → [/] dans l' apprentissage [///] non / dans la → +

*ENT: en apprentissage ? ou → dans l' enseignement ?

*PSU: oui // dans@[108]g l'@g enseignement@g // parce que / bon je crois que → / c' est plus → [/] plus sûr ?

[C2W02]

En una conversación, como ya hemos expresado, es habitual que no sólo ocurran interrupciones por parte de los hablantes, sino también solapamientos, como veremos más adelante, o que se construya el enunciado de forma cooperativa. Puede ocurrir así que un hablante comience un enunciado, sea interrumpido por el oyente para expresar alguna matización y después continúe. O bien que el oyente participe en la construcción del enunciado, por ejemplo, si el hablante olvida una palabra, se equivoca o bien pide la ayuda del interlocutor. En los corpus de aprendientes, este tipo de marca es bastante habitual, por lo que

[108] La etiqueta @ se describe en el epígrafe 4.1.2.4 del presente capítulo, y hace referencia a todos los fenómenos relacionados con la interlengua del aprendiente, o a los fenómenos propios de la lengua oral.

siguiendo las indicaciones ya implantadas en el corpus de aprendientes de ELE del LLI-UAM, convinimos en utilizarla. Esta marca de continuación se expresa a través del signo ¬ justo después del código de tres letras que muestra el inicio de turno y antes del enunciado que el hablante pretende continuar:

*ALB: en comparaison avec → &eh / l' anglais // <parce que → > /

*ENT: [<] <l' anglais> ? ///

*ALB: ¬ je crois que le &fran [/] &eh [/] le [/] le &f [/] le français c' est / hhh {%act: blow} / plus courtisan <et → >

*ENT: [<] <aha!> ///

*ALB: ¬ plus → / lyrique // <et → > [///]

[C1M01]

4.1.2.3 Fenómenos relacionados con el habla espontánea

4.1.2.3.1 Repeticiones y reinicios o reformulaciones

Utilizamos la marca [/] cuando encontramos la repetición de una palabra o secuencia de palabras cortas, situándose entre los elementos que se repiten:

*ENT: et vous pensez / que le français / c' est → difficile // c' est facile → // c' est → / moins compliqué que d'autres langues → ?

*VIC: hhh {%act: doubt} / pour moi c' est → [/] c' est bien ///

[A2M03]

Las reformulaciones o reinicios reflejan generalmente las dudas del hablante para expresar algo. Esta reformulación puede influir sólo a una palabra o un fragmento de sintagma o bien, a una parte mayor del enunciado, provocando un reinicio completo o la reconstrucción sintáctica del enunciado para continuar con la misma idea, pero con distintas palabras (algo frecuente en aprendientes que ya tienen algo de dominio de la lengua, que son capaces de poner en marcha esta estrategia

de comunicación para salvar ciertas dificultades de la lengua meta o de su producción). En el caso de un reinicio sin cambio sintáctico, utilizaremos la misma marca que para las repeticiones [/]:

*ABE: hhh {%act: blow} / hhh {%act: laugh} /// &eh / c'est très bien passée et → // je suis arrivée en → / septembre // et → / j'habitais un → [/] dans une → / cité universitaire /// &eh / à côté de la fac@oral ///

[B2W01]

En el caso de reformulación sintáctica, se establece el uso de la marca [///]:

*ARI: quel est ce que → [///] quel genre de film / tu aimes [/] tu aimes → / voir ?

*ANM: j' aime → le cinéma de [/] d' auteur ? [/] <auteur> ///

[A2W01]

4.1.2.3.2 Autointerrupciones o abandonos del mensaje

Este tipo de fenómeno aparece cuando el hablante interrumpe bruscamente lo que está diciendo para comenzar algo que expresa otra idea distinta o que nada tiene que ver con el enunciado precedente. Se diferencia de la reformulación sintáctica en que no sólo cambia la forma, sino el contenido, puesto que en este caso se inicia una nueva idea. La marca que lo expresa es el signo = :

*PEP: et je → / cherche / &eh [/] sur → Internet // &eh [/] &nim [/] n' importe → quelle → &ch [/] quel → groupe ou quelle chanson // &eh / français / pour connaître → = <j' ai connu → >

*ENT: [<] <ah! ben c' est bien> ///

*PEP: ¬ dans le réseau // quelques / groupes de → [/] de → &a [/] de rap //

[C2W01]

4.1.2.3.3 Alargamientos

En la interacción es frecuente que los hablantes alarguen ciertas vocales o finales de palabras con la intención de ganar tiempo o reflexionar acerca de lo que va a decir. Este fenómeno se marca junto a la palabra en la que se produce con el símbolo → :

*CAU: oui /// &mm / &eh / maintenant je ne travaille pas /// parce que → je me prépare / pour → le concours // d' éducation secondaire ///

[B1W02]

4.1.2.3.4 Palabras incompletas o fragmentos de palabras

A lo largo del discurso es muy habitual que haya palabras que no lleguen a pronunciarse o como hemos visto anteriormente, que se reformulan para mejorar la expresión o el significado. Este efecto produce palabras incompletas o fragmentos de palabras que suelen acompañar a repeticiones o reformulaciones, y que se marcan en la transcripción con el signo &. Es habitual cuando los aprendientes se autocorrigen, cuando dudan o si intentan mejorar su expresión:

*TAG: oui peut-être la → / prochaine année // &eh / parce que / je crois que → / j' aurais plus / d' &inde [/] d' indépendance → et → / je pourrais / faire les choses / quand → / je → voudrais ///

[B2W03]

4.1.2.3.5 Apoyos vocálicos

Los apoyos vocálicos son vocalizaciones que el hablante suele emplear para mantener el turno de palabra, y que le permiten reflexionar o estructurar su discurso. No suelen tener ningún significado fuera del aspecto pragmático, y en nuestro corpus se señala con el carácter & (para

que no sea contado en el total de palabras) y la interjección correspondiente, fundamentalmente ***&ah, &eh, &euh*** y ***&mm***.

*LUI: hhh {%act: click} / d' accord /// moi je suis → / &eh / un étudiant → / de langue étrangère française / &eh / ici à la fac de → / Cxxx Rxxx ///

[B2M01]

4.1.2.3.6 Segmentos incomprensibles

Entendemos por segmentos incomprensibles aquellas partes del enunciado que no pueden descifrarse, por problemas de sonido o bien, por una mala producción del locutor. En estos casos utilizamos la marca xxx[109]:

*JOA: &eh / j' ai visité → / Paris → / l' an → dernier → /// et il faut → / déjà → / plus xxx [/] j' ai visité → / Niza@spa / Nice → ///

[C2M02]

4.1.2.3.7 Solapamientos

En la interacción oral es muy frecuente, que al igual que exista una negociación de los turnos o una construcción conjunta del enunciado, también existan algunas partes donde los participantes hablen a la vez, produciendo simultáneamente su enunciado. Ocurre principalmente cuando el oyente interrumpe al hablante, y en nuestro corpus se marca poniendo el enunciado simultáneo entre los signos < > , y añadiendo la marca [<] al comienzo del siguiente turno, que se solapa con el anterior

[109] Utilizamos también la marca xxx en aquellas partes del discurso que pretendemos anonimizar, sobre todo, en relación con los datos personales del aprendiente. En estos casos, entendemos lo que quiere decir, pero lo omitimos conscientemente. Es por ello que se señala en el texto con la marca xxx pero precedida de la letra inicial de la palabra, generalmente, nombres propios. Por ejemplo, con Sxxx para Sara o Vxxx para Valencia (ejemplos no presentes en nuestro corpus).

(aunque es posible que los solapamientos se produzcan entre varias intervenciones, con lo que pondríamos la marca al principio de todos aquellos que abarque):

*ENT: et aimerais tu / retourner en France pour vivre ou pour travailler là-bas → ? <ou → > ...

*ABE: [<] <oui> / &eh / l' année prochain je vais → / demander le bourse / pour → / être → / auxiliaire de → conversation // mais je ne sais pas si je vais → / le {%alt: de} demander pour aller en France / ou en Belgique ///

*ENT: aha! /// et si → / &eh [/] &s [///] bon et pour aller en France / tu irais où → ? exactement /// bon / &eh / <peu importe>

*ABE: [<] <je ne sais pas encore> ///

[B2W01]

4.1.2.3.8 Signos paralingüísticos

En una conversación es muy frecuente que además de palabras u oraciones (signos lingüísticos) existan otra serie de elementos o recursos como gestos, sonidos onomatopéyicos o no reconocidos, que además de contribuir a la riqueza expresiva de la interacción, aportan un significado al contenido, complementándolo. Los signos paralingüísticos se marcan en el texto transcrito como hhh, seguidos de la explicación entre llaves y unida a la etiqueta %act, que introduce la aclaración o comentario. Así, obtendríamos una marca de la siguiente forma: *hhh {%act: explicación del signo}*. Entre los más frecuentes encontramos:

assent [asentimiento]
blow [soplido]
click [chasquido]
cough [tos]
doubt [duda]
clearing of throat [carraspeo]
question [pregunta]
sigh [suspiro]

breathe [respiración]
onomatopeia [onomatopeya]
laugh [risa]
inhalation [inspiración]

*CAU: &eh / je pense / que → Axxx / &eh / c' est une grand village // &eh / qui est → / très très bien / parce que → / &mm / hhh {%act: click} / ici → a / beaucoup de choses // &eh / &mm / dont j' ai besoin /// hhh {%act: inhalation} / mais → / &eh / en plus / Axxx / est → / &mm [/] est → / très → / cerca@spa ?

*ENT: très proche ///

*CAU: très@g proche@g / &eh / de Madrid ///

*ENT: hhh {%act: assent} ///

[B1W02]

4.1.2.9 Comentarios

Además de los ya expresados con las etiquetas *%alt* (desviaciones o alteraciones en la producción) y *%act* (definición de los signos paralingüísticos), utilizamos la marca *%com*, para incluir comentarios que el transcriptor o revisor considere oportunos para una mejor comprensión de la interacción oral. En nuestro corpus se suele utilizar para comentar aspectos del contexto de la interacción, corregir algunos de los errores en formas conjugadas o reseñar errores fonéticos, realizar comentarios sobre problemas eventuales del entorno de grabación o para explicar algunos detalles culturales que el lector puede no conocer.

*JCA: ¬ [<] <j' ai besoin> de → / avoir / à mon côté → / &eh / le Petit Robert {%com: french monolingual dictionnary} ///

[B2M02]

*VIC: le → [/] le moyen / s' appelle / Vxxx /// il a → / de / vingt trois ans // et il étude@c {%com: il étudie} ///

*ENT: hhh {%act: assent} ///

[A2M03]

4.1.2.10 Comentarios sobre fallos o alteraciones de producción oral

Entre los comentarios que podemos hacer del enunciado también encontramos los indicados con la marca *%alt*. En este caso nos referimos a alteraciones en la forma o en la producción oral de una palabra, como la elisión de un sonido o la sustitución de un sonido por otro.

Este tipo de comentario plantea algunas dificultades, puesto que existen ciertos casos en los que no sabemos si es un fallo de producción, o si se trata de un error, en los que el hablante desconoce el léxico. En el caso de que consideremos que se trata de un error, utilizaremos otro tipo de etiqueta descrita más adelante como es la de creación léxica.

*TAG: bon / dès → que j' étude@c {%com: j'ai étudié/ j'étudiais} / au lycée // j' ai aimé → / les langues // surtout le français /// mais je ne peux pas exprimer pourquoi → // j' aime sa → sonorité / c' est une langue → / très belle {%alt: belt} [/] belle // et aussi → / quand → j' ai connu la littérature // je l' ai aimée beaucoup // et aussi je → [/] j' aime étudier de l' anglais // parce que c' est très important / mais → / je préfère la → [/] la philologie française ///

[B2W03]

4.1.2.11 Pausas prolongadas

Con el signo # marcamos aquellas pausas largas en los que el hablante no produce ningún enunciado, y que generalmente denotan tiempo para reflexionar ante dudas o al no recordar un término, y que permiten al hablante estructurar el nuevo enunciado:

*ANM: et → / hhh {%act: blow a raspberry} / &eh # &eh / je → [/] j' ai eu une → époque / que j' ai → écouté → / &mm / Daft Punk // je crois que c' est <française> +

*ARI: <oui / c' est français> /// hhh {%act: assent} ///

[A2W01]

4.1.2.12 Interjecciones

Las interjecciones son muy habituales en la interacción oral, poseen un significado pragmático, y generalmente se alude a ellas como la materialización a través de la lengua de determinados sentimientos del hablante. Cada lengua posee una lista de interjecciones frecuentes y para el francés, recogemos aquellas expresadas en las normas de transcripción del corpus C-ORAL-ROM en su versión francesa (Équipe DELIC, 2004). Las interjecciones aparecen en nuestro corpus acompañadas del signo de exclamación de cierre (!), sin espacios (para diferenciarlas de los enunciados exclamativos, en los que añadimos un espacio al final del mismo y antes del signo de exclamación). Podemos ver un ejemplo en:

*REY: ¬ et pour / améliorer la participation à / un procès de concours → / <&eh> /

*ENT: [<] <ah! ben c' est> <bien // c' est très [/] c' est intéressant> ///

[C1W03]

4.1.2.4. Etiquetas para el análisis de la interlengua

Nuestro objetivo es compilar un corpus donde lo más importante sea la interlengua del aprendiente. Para ello y con el fin de que podamos tener presentes en el análisis algunos fenómenos reseñables, utilizaremos la etiqueta @ seguida de una letra al final de la palabra o expresión afectada[110] para mostrar aquellos aspectos que queremos controlar. Entre ellos podemos destacar los siguientes:

[110] Esta etiqueta se encuentra presente en algunos de los corpus de aprendientes más importantes del francés, como es el caso FLLOC, del que recogemos esta norma de transcripción.

- **Extranjerismos** o palabras de otras lenguas que se intercalan en el discurso de forma inconsciente, normalmente porque el uso de esa palabra está ya muy consolidado en la lengua meta, o bien porque existe un conocimiento mayor de esa lengua, interfiriendo así con el enunciado en momentos de duda o nerviosismo. Es muy frecuente que ocurra con la lengua materna del hablante y con el inglés. Marcamos este fenómeno al final de la palabra con el símbolo @ seguido del código de tres letras correspondiente al nombre de la lengua en inglés. Las marcas de extranjerismos más frecuentes de nuestro corpus son las siguientes: @spa (español), @eng (inglés) y @ita (italiano).

*LOU: ¬ &eh / &mm / aller au cinéma → // hhh {%act: click} / &eh / aller à → / faire de shopping@eng ///

[B1W03]

*VIC: &eh / je → travaille [/] je suis emploi de → / l' administración@spa → ///

[A2M03]

- **Creación léxica** por parte del aprendiente, inventando palabras que no existen en la lengua meta, bien porque desconocen su forma exacta y ello les induce a error, o lo más común, porque tratan de asimilar la forma de la palabra de la lengua materna a la lengua meta, traduciéndola casi literalmente. Se marca con la etiqueta @n.

*ALR: &eh / je pense → &eh / faire un entreviste@n / &eh / pour les → [/] pour les professeurs ///

[A1M01]

- **Conjugación errónea de una forma verbal**, que se expresa mediante la marca @c al final de aquellas palabras que contienen verbos cuya conjugación no es correcta. A continuación, se expresa como comentario la forma verbal que sería adecuada

según la norma para dicho contexto[111]. Generalmente ocurre debido a hipergeneralizaciones de las reglas conocidas por el aprendiente.

*SUM: &eh / mes &pa [/] mes parents / &fer [/] &fai [/] faisent@c {%com: font} / du [/] font / hhh {%act: laugh} [///] mes parents → / font → du &ca [/] du camping //

[A2W02]

- **Imitación o repetición de las palabras del interlocutor:** El aprendiente a lo largo de su discurso es posible repita o versione algunas de las expresiones, formas lingüísticas o palabras que produzca su interlocutor (habitualmente, el/la entrevistador/a). Esta imitación o repetición puede aparecer por falta de fluidez o creatividad, escogiendo una forma ya aparecida y validada por el interlocutor, o bien porque desconozca la palabra o expresión y el interlocutor se la haya facilitado. Este fenómeno se marca con el uso de @g al final de la palabra[112]. Además, sólo marcamos la primera aparición, ya que si se repite de nuevo a lo largo del discurso de forma adecuada, entendemos que el hablante la ha adaptado con efectividad.

*STE: l' anglais /// tu l' as appris comme le français à l' école / <ou → > ?

*ANT: [<] <non> // je l' ai@g appris@g à@g → [/] au lycée /// et → [/] et quelques → [/] quelques → / cours //

[C1M02]

[111] Hemos de señalar que sólo atendemos a errores de la conjugación relativos a la forma. En el caso de errores derivados de una inadecuación del tiempo conjugado al contexto, o bien, problemas frecuentes como un uso inadecuado del auxiliar en verbos compuestos como el *passé composé*, no lo señalamos como tal, pero sí se tiene en cuenta para el análisis de errores posterior.

[112] Tomando las pautas establecidas en los corpus de aprendientes de Southampton FLLOC y SPLLOC.

4.1.2.6 Etiquetas especiales que denotan fenómenos de la gramática de la lengua oral

Como venimos comentando a lo largo de nuestro estudio, nos interesamos especialmente por el grado de asimilación del uso de la lengua oral en nuestros aprendientes. Por consiguiente, y con el fin de poder analizar de forma objetiva el uso de la llamada "gramática de la lengua oral" en sus producciones, nos propusimos introducir una nueva marca que resultase de ayuda a la hora de analizar la interlengua. En este caso etiquetamos aquellos fenómenos de la lengua oral reseñables con dos tipos de marcas: *@oral* y *%oral.*

La primera, *@oral*, acompaña a palabras aisladas que suelen ser propias del registro oral, como puede ser el caso de *fac, boulot, sympa*… que si bien son palabras plenamente aceptadas y enormemente utilizadas en la conversación corriente, suelen ser aún concebidas como parte del registro más familiar o conversacional.

*ABE: bonjour /// hhh {%act: laugh} /// je suis Axxx /// &eh / j' habite à &Ma [/] à Mxxx // qui est un petit {&alt: petip} → / village / &eh / d' Axxx /// &mm / j' étude@c {%com: j'étudie} à → / la fac**@oral** de → / lettres // à Cxxx Rxxx ///

[B2W01]

*PEP: [<] <ça@oral c' est → [/] aussi> = oui /// et aussi &l [/] l' utilisation de → [/] des subordonnées → {%alt: sobordonnées} et les → prépositions qu' on**@oral** doit utiliser // <pour placer → aussi> ///

[C2W01]

La segunda, *%oral*, se considera un tipo de comentario, por lo que aparece entre llaves, y está destinada a explicar o comentar fenómenos sintácticos más complejos, que abarcan en ocasiones más de una única palabra y en los que sería complicado utilizar de forma clara la primera etiqueta. Con esta marca solemos aludir a fenómenos de la lengua oral considerados como pertenecientes a la gramática propia de esta variedad como la *dislocation à gauche*, la ausencia de la partícula *'ne'* en la forma negativa, la contracción de sujeto y verbo en la pronunciación como en

't'as bien aimé?' o incluso, la desaparición del sujeto en fórmulas frecuentes como '*je ne sais pas'* (lo que da lugar a '*sais pas*').

*STE: donc pour toi / c'est qu' y a **{%oral**: 'il' absent} / de plus difficile // dans le français ///

[B2M02]

*NOE: c' est sympa**@oral** mais → / bon j' ai pas **{%oral**: 'ne' absent} / pu → [/] j' ai pas pu {%oral: 'ne' absent} trop profiter parce que → / je travaille aussi alors → / jusqu' à présent → / j' ai / à peine pu → / &eh / aller en cours → // c' est cette année / où je commence / aller en cours → // bon c' est vrai c' est → [/] c' est une très bonne expérience ///

[C1W02]

Este tipo de fórmulas ha sido considerado normalmente como un error en el uso del registro, relacionado con esa visión de la lengua donde la variedad estándar (y normativa) tiene mucho que ver con la variedad escrita. En nuestro caso, consideramos su uso como un elemento positivo, ya que denota un grado de competencia mayor por parte del aprendiente. Es evidente que un hablante que no sea posea un cierto grado de fluidez y competencia no será capaz de captar este tipo de matices. De hecho, en nuestro corpus es mucho más frecuente en aquellos aprendientes de niveles superiores y en aquellos que han realizado estancias prolongadas en países de habla francófona.

4.1.2.7 Resumen de los criterios de transcripción

	MARCA	EJEMPLO
Turnos	***(Código de tres letras en mayúscula): *ABC:**	*ENT: hhh {%act: assent} / tu travailles // et → qu' est-ce que tu fais comme → / travail ? *ALR: &eh / je suis professeur / &eh / de [/] dans l'école ///

Pausas entonativas		**/** **//**	*ARI: hhh {%act: assent} /// tu m'as dit que tu → / écoutes de la musique // quel genre de musique / tu aimes ?
Marcas de continuación de turno y final de enunciado	**Enunciado interrogativo**	**?**	*ENT: et Sxxx / vous → [/] vous → êtes née ici ? *SUB: non je suis née à → / Paris ///
	Enunciado exclamativo	**!**	*SUB: hhh {%act: click} / elle@g étudie@g / des ? mathématiques et ? [/] et ? &ingé [/] ingénieur informatique / les <deux> ? + *ENT: [<] <ah! les> <deux> !
	Enunciado asertivo	**///**	*SAN: &eh / parce que → / &esa → [/] &eh / il est → / &ah [/] intéressant pour moi ///
	Enunciado inacabado o suspendido	**…**	*STE: [<] <hhh {%act: laugh}> /// donc / &eh / si tu → [/] si tu pouvais me décrire un lieu en France que → [/] que tu as visité / que tu as aimé / <ou → > …
	Interrupción por otro hablante	**+**	*PSU: ah! oui / voilà /// mais → / hhh {%act: click} / finalement → / je → [/] j'ai décidé de → [/] de suivre / dans → [/] dans l' apprentissage [///] non / dans la → + *ENT: en apprentissage ? ou → dans l'enseignement ?
	Continuación de turno	**¬**	*ALB: ¬ je crois que le &fran [/] &eh [/] le [/] le &f [/] le français c' est / hhh {%act: blow} / plus courtisan <et → > *ENT: [<] <aha!> /// *ALB: ¬ plus → / lyrique // <et → > [///]
Repeticiones		**[/]**	*ENT: et vous pensez / que le français / c' est → difficile // c' est facile → // c' est → / moins compliqué que d'autres langues → ? *VIC: hhh {%act: doubt} / pour moi c'

			est → [/] c' est bien ///
Reinicios	**Reinicio o reformulación**	**[/]**	*ABE: hhh {%act: blow} / hhh {%act: laugh} /// &eh / c'est très bien passée et → // je suis arrivée en → / septembre // et → / j'habitais un → [/] dans une → / cité universitaire /// &eh / à côté de la fac@oral ///
	Reformulación sintáctica	**[///]**	*ARI: quel est ce que → [///] quel genre de film / tu aimes [/] tu aimes → / voir ?
Autointerrupción		**=**	*PEP: et je → / cherche / &eh [/] sur → Internet // &eh [/] &nim [/] n' importe → quelle → &ch [/] quel → groupe ou quelle chanson // &eh / français / pour connaître → = <j' ai connu → > *ENT: [<] <ah! ben c' est bien> /// *PEP: ¬ dans le réseau // quelques / groupes de → [/] de → &a [/] de rap //
Alargamiento		**→**	*CAU: oui /// &mm / &eh / maintenant je ne travaille pas /// parce que → je me prépare / pour → le concours // d' éducation secondaire ///
Fragmentos de palabras		**&**	*TAG: oui peut-être la → / prochaine année // &eh / parce que / je crois que → / j' aurais plus / d' &inde [/] d' indépendance → et → / je pourrais / faire les choses / quand → / je → voudrais ///
Apoyos vocálicos		**&ah &eh &mm &euh**	*LUI: hhh {%act: click} / d' accord /// moi je suis → / &eh / un étudiant → / de langue étrangère française / &eh / ici à la fac@oral de → / Cxxx Rxxx ///
Segmentos incomprensibles		**xxx**	*JOA: &eh / j' ai visité → / Paris → / l' an → dernier → /// et il faut → / déjà → / plus xxx [/] j' ai visité → /

		Niza@spa [/] Nice → ///
Solapamientos	**< >** **[<]**	*ENT: et aimerais tu / retourner en France pour vivre ou pour travailler là-bas → ? <ou → > ... *ABE: [<] <oui> / &eh / l' année prochain je vais → / demander le bourse / pour → / être → / auxiliaire de → conversation // mais je ne sais pas si je vais → / le {%alt: de} demander pour aller en France / ou en Belgique /// *ENT: aha! /// et si → / &eh [/] &s [///] bon et pour aller en France / tu irais où → ? exactement /// bon / &eh / <peu importe> *ABE: [<] <je ne sais pas encore> ///
Signos paralingüísticos	**hhh {%act : x}**	*CAU: &eh / je pense / que → Axxx / &eh / c' est une grand village // &eh / qui est → / très très bien / parce que → / &mm / hhh {%act: click} / ici → a / beaucoup de choses // &eh / &mm / dont j' ai besoin /// hhh {%act: inhalation} / mais → / &eh / en plus / Axxx / est → / &mm [/] est → / très → / cerca@spa ? *ENT: très proche ///
Fallos o alteraciones de la producción oral	**{%alt}**	*TAG: bon / dès → que j' étude@c {%com: j'ai étudié/ j'étudiais} / au lycée // j' ai aimé → / les langues // surtout le français /// mais je ne peux pas exprimer pourquoi → // j' aime sa → sonorité / c' est une langue → / très belle {%alt: belt} [/] belle // et aussi → / quand → j' ai connu la littérature // je l' ai aimée beaucoup // et aussi je → [/] j' aime étudier de l' anglais // parce que c' est très important / mais → / je préfère la → [/] la philologie française ///

Comentarios		**%com**	*JCA: ¬ [<] <j' ai besoin> de → / avoir / à mon côté → / &eh / le Petit Robert {%com: French monolingual dictionnary} ///
Pausas prolongadas		**#**	*ANM: et → / hhh {%act: blow a raspberry} / &eh # &eh / je → [/] j' ai eu une → époque / que j' ai → écouté → / &mm / Daft Punk // je crois que c' est <française> + *ARI: <oui / c' est français> /// hhh {%act: assent} ///
Interjecciones		**Ah! Eh! Bof!**	*ENT: [<] <ah! ben c' est> <bien // c' est très [/] c' est intéressant> ///
Análisis de la Interlengua	**Extranjerismos**	**@código lengua i.e : @spa**	*LOU: ¬ &eh / &mm / aller au cinéma → // hhh {%act: click} / &eh / aller à → / faire de shopping@eng ///
	Creación léxica	**@n**	*ALR: &eh / je pense → &eh / faire un entreviste@n / &eh / pour les → [/] pour les professeurs ///
	Conjugación errónea	**@c**	*SUM: &eh / mes &pa [/] mes parents / &fer [/] &fai [/] faisent@c {%com: font} / du [/] font / hhh {%act: laugh} [///] mes parents → / font → du &ca [/] du camping //
	Repetición de palabras del interlocutor	**@g**	*STE: l' anglais /// tu l' as appris comme le français à l' école / <ou → > ? *ANT: [<] <non> // je l' ai@g appris@g à@g → [/] au lycée /// et → [/] et quelques → [/] quelques → / cours //
	Palabras o fenómenos propios de la lengua oral	**@oral** **{%oral:}**	*NOE: c' est sympa@oral mais → / bon j' ai pas {%oral: 'ne' absent} / pu → [/] j' ai pas pu {%oral: 'ne' absent} trop profiter parce que → / je travaille aussi alors → / jusqu' à présent → / j' ai / à peine pu → / &eh

			/ aller en cours → // c' est cette année / où je commence / aller en cours → // bon c' est vrai c' est → [/] c' est une très bonne expérience ///

Tabla 9: Resumen de las convenciones de transcripción del corpus CORAF.

5. Alineamiento

Tras finalizar el proceso de transcripción se procede a las tareas de al alineamiento, consistente en la sincronización del texto que hemos transcrito con la parte del enunciado que le corresponde en el archivo de audio. El alineamiento puede realizarse siguiendo distintas convenciones y puede incluir turnos completos de cada hablante o bien, enunciados. En CORAF mantenemos las convenciones establecidas en los corpus orales del LLI-UAM, alineando cada transcripción por enunciados acabados o completos.

El alineamiento se realiza, como ya hemos comentado, por medio del programa utilizado para la transcripción, Transana, que permite hacerlo de forma simultánea a esta. Para ello utilizamos uno de los comandos del programa (Control + T), que inserta una marca de tiempo en milisegundos en aquellas partes donde el transcriptor (o revisor) ha establecido.

En CORAF hemos considerado alinear nuestras transcripciones por enunciados, lo que supone introducir la marca temporal tras las marcas de final de enunciado, enunciado suspendido e interrupción del hablante (representadas por ///, !, ?, … , +) y en los solapamientos de hablantes, tras la marca de < >.

Una vez realizado el alineamiento obtenemos una transcripción mucho más rica, puesto que programas de visualización de datos como Transana nos permiten escuchar los archivos de audio y señalar, al mismo tiempo, el segmento del texto al que pertenece (de forma similar a los subtítulos de una película), lo cual facilita la comprensión del documento auténtico y la hace mucho más accesible a todo tipo de

usuarios, como por ejemplo, alumnos en instrucción de segundas lenguas o futuros docentes en prácticas de uso de corpus directo.

6. Tratamientos posteriores del corpus: estandarización y conversión a XML

Uno de los procesos finales en la compilación del corpus es dotar al mismo de una estructura y formato común, al igual que haríamos con un texto o con una colección de datos concreta.

Así, nuestra intención, como ya expresamos en el diseño de CORAF, es proporcionar nuestro corpus siguiendo, dentro de lo posible, los estándares existentes, o las características técnicas que puedan facilitar su reutilización en otras investigaciones o su uso con herramientas y aplicaciones informáticas especiales para corpus.

Para ello, organizamos nuestro corpus con una estructura sencilla, utilizando una carpeta general que contiene en subcarpetas, todas las transcripciones y archivos de audio relacionados. Estas subcarpetas se nombran y organizan en torno a los niveles del MCER representados en el corpus (A1, A2, B1, B2, C1, C2).

Dentro de estas carpetas, encontraremos los archivos en audio en formato .wav, y sus correspondientes transcripciones en formato de texto plano (txt), siguiendo una codificación de UTF-8, que garantiza que no existan inconsistencias a la hora de usarlos en distintos programas y aplicaciones informáticas.

No obstante, las últimas investigaciones en LC, sitúan, además del formato específico en TEI (*Text Encoded Initiative*), al formato XML (*eXtended Markup Language*) como uno de los más recomendables a la hora de difundir los archivos de los corpus. Por tanto, nuestra intención es ofrecer los archivos de texto también en formato XML.

Transana es un programa muy completo de edición, y permite exportar los archivos de texto creados, pero sólo en formato RTF. Aún así, es posible realizar la exportación del texto seleccionándolo, cortando y pegando hacia otros programas como *Microsoft Word* o *EditPlus*, sin que añada ningún tipo de marcas adicionales que puedan perjudicar al fichero.

En el caso del formato XML es algo más complicado, ya que no existen programas a nuestro alcance que realicen esta función con tan sólo exportarlo. Por tanto, para la conversión a XML se utilizará un programa especialmente creado en el LLI-UAM que facilita el proceso realizándolo de forma automática sobre todos los ficheros, que, además, les dota del formato establecido para todos los corpus del laboratorio (similar al del corpus oral C-ORAL-ROM).

Una vez creados todos los ficheros en txt y XML nuestro corpus podrá utilizarse en los estudios, análisis o aplicaciones que se deseen, así como ser sucesivamente enriquecido añadiendo distintos tipos de anotación.

7. CORAF: estructura final

CORAF se compone en total de 30 entrevistas a aprendientes de los seis niveles del MCER y consta de 61.092 palabras, distribuidas en 33.915, correspondientes a aprendientes, y 27.177 a los/las entrevistadores/as. Así, abarca un total de 7 horas y 22 minutos, que se reparten en entrevistas con una duración media de 14 minutos y 45 segundos.

El conjunto de los archivos de nuestro corpus posee así las siguientes características:

ARCHIVO	NIVEL MCER	DURACIÓN ENTREVISTA	Nº TOTAL PALABRAS	Nº PALABRAS APRENDIENTE	PALABRAS ENT.	TURNOS	Nº FENÓMENOS ORAL
A1M01	A1	0:10:44	1371	445	926	189	0
A1M02	A1	0:12:03	1639	288	1351	212	1
A1W01	A1	0:13:27	1196	429	767	162	0
A1W02	A1	0:12:54	1528	741	787	214	0

ARCHIVO	NIVEL MCER	DURACIÓN ENTREVISTA	Nº TOTAL PALABRAS	Nº PALABRAS APRENDIENTE	PALABRAS ENT.	TURNOS	Nº FENÓMENOS ORAL
A1W03	A1	0:11:16	1255	603	652	162	3
	A1	**1:00:24**	**8503**	**4110**	**4393**	**939**	**4**
A2M01	A2	00:12:49	1648	605	1043	266	2
A2M02	A2	00:13:43	2189	1180	1009	257	37
A2M03	A2	00:10:18	1460	485	975	215	1
A2W01	A2	00:13:54	1444	846	598	212	1
A2W02	A2	00:14:38	1762	994	768	361	2
	A2	**1:05:22**	**8503**	**4110**	**4393**	**1311**	**43**
B1M01	B1	00:12:57	1688	749	939	175	6
B1W01	B1	00:15:23	2062	1011	1051	319	2
B1W02	B1	00:17:08	2173	1111	1062	363	12
B1W03	B1	00:14:54	2262	1026	1236	313	27
B1W04	B1	00:13:57	1514	1011	503	135	1
	B1	**1:14:19**	**9699**	**4908**	**4791**	**1305**	**48**
B2M01	B2	00:18:59	2695	1676	1019	263	46
B2M02	B2	00:15:07	1794	1081	713	215	1
B2W01	B2	00:12:57	1975	904	1071	160	14
B2W02	B2	00:16:56	2392	1780	612	126	24
B2W03	B2	00:15:47	2423	1417	1006	262	14

ARCHIVO	NIVEL MCER	DURACIÓN ENTREVISTA	Nº TOTAL PALABRAS	Nº PALABRAS APRENDIENTE	PALABRAS ENT.	TURNOS	Nº FENÓMENOS ORAL
	B2	**1:19:46**	**11279**	**6858**	**4421**	**1026**	**99**
C1M01	C1	00:16:59	2416	1464	952	298	11
C1M02	C1	00:13:24	1845	1095	750	228	47
C1W01	C1	00:17:35	2841	1825	1016	325	23
C1W02	C1	00:14:11	2694	1875	819	179	76
C1W03	C1	00:18:19	2569	1608	961	309	16
	C1	**1:20:28**	**12365**	**7867**	**4498**	**1339**	**173**
C2M01	C2	00:14:16	1750	1162	588	189	5
C2M02	C2	00:15:54	2348	1518	830	330	23
C2W01	C2	00:20:00	2762	1756	1006	390	26
C2W02	C2	00:14:40	2564	1487	1077	287	24
C2W03	C2	00:17:14	2833	1743	1090	207	50
	C2	**1:22:04**	**12257**	**7666**	**4591**	**1403**	**128**
TOTAL		7:22:23	61092	33915	27177	7323	495

Tabla 10: Resumen de los datos representativos del corpus oral de aprendientes CORAF.

De estos datos se desprende que la complejidad del discurso va en aumento en función del nivel, pese a que en el nivel C2 se invierte la tendencia, debido, sin duda, a las características específicas de dichos aprendientes. Podemos observar distintos fenómenos con la ayuda de las siguientes gráficas:

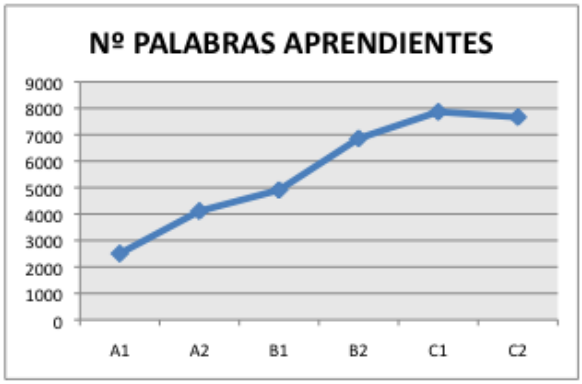

Gráfico 9: Número total de palabras producidas por los aprendientes de cada nivel del MCER.

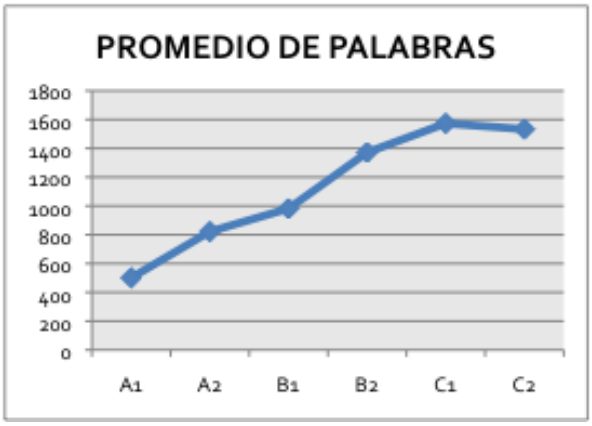

Gráfico 10: Promedio de palabras de los aprendientes según el nivel del MCER representado en CORAF.

El promedio de palabras producidas por aprendientes de nivel A1 es 501,2; para A2 aumenta a 822; sigue creciendo para el nivel B1 con 981,6 palabras y para B2 con 1.371,6. Los niveles C1 y C2, aumentan con

respecto al nivel independiente (nivel B), siendo para C1, 1.573,4 y para C2, invirtiendo la tendencia, 1.533,2 palabras.

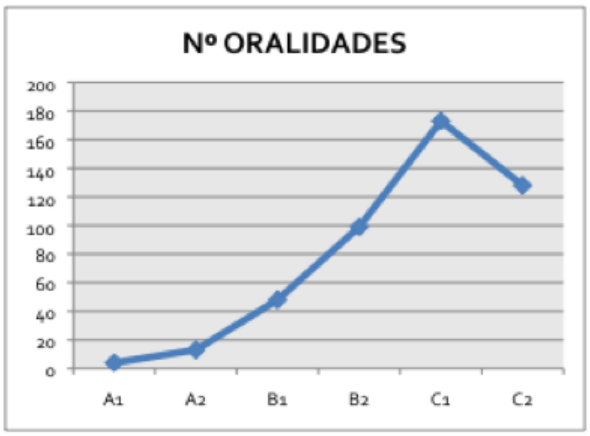

Gráfico 11: Número de fenómenos de oralidad presentes en cada nivel del MCER representado en CORAF

Así, vemos cómo los aprendientes no sólo van utilizando cada vez más palabras, sino que también suelen incluir más fenómenos de la lengua oral. No obstante, los fenómenos de la lengua oral no dependen tanto del nivel del MCER que ostenten, sino de la experiencia lingüística previa en ambiente endógeno, ya que coincide que los hablantes con un mayor uso de fenómenos propios de la lengua oral son aquellos que más tiempo han pasado en países francófonos.

8. Conclusiones

A lo largo de este capítulo hemos explicado cómo se ha ido desarrollando la compilación e implementación del corpus oral CORAF, desgranando cada una de las fases necesarias para su consecución, empezando por la grabación, y siguiendo con el tratamiento del audio y la digitalización, la transcripción y el alineamiento, y finalmente, la estandarización y conversión de los archivos de audio y texto a otros formatos.

En todas las fases hemos descrito aquellos aspectos qué hemos tenido en cuenta en función del diseño previo de nuestro corpus, incidiendo en aquellos puntos en que las circunstancias de la grabación nos han obligado a modificar el diseño, como ocurre con el número de aprendientes disponible para cada nivel del MCER. Creemos necesario desgranar todas las dificultades encontradas, ya que resultará beneficioso para lograr una metodología más efectiva en el futuro y nos ayudará en posteriores ampliaciones de nuestro corpus.

Evidentemente, la fase de recogida de datos será decisiva para la conformación del corpus y para los resultados que podamos obtener de nuestros análisis. Así, podemos afirmar que es muy frecuente que el diseño inicial se vea sustancialmente modificado, sobre todo ante la limitación que el tiempo y el número de datos obtenidos suele imponer.

Así las cosas, el hecho de no poder contar en CORAF con tantos entornos de grabación como habíamos previsto, además de un menor número de aprendientes en aquellas EOI que sí aceptaron participar en nuestro proyecto, unido, además, a la limitación de tiempo existente, merma la representatividad de nuestro corpus. Sin embargo, y pese a todos los inconvenientes, el hecho de conseguir reunir algo más de 33.000 palabras (exclusivamente de aprendientes) nos induce a pensar que podremos establecer la base de hipótesis sólidas sobre la interlengua verificables en un futuro con sucesivas ampliaciones del número de aprendientes por nivel. No en vano, la mayoría de los corpus específicos o de especialidad, como hemos comentado en capítulos anteriores, se limita las 30.000 palabras totales[113] en el entorno académico. Por tanto, nuestro corpus no dista tanto de otros de su entorno más inmediato, con los que podremos establecer comparaciones y llevar a cabo diferentes análisis de la interlengua (sobre todo en análisis contrastivos de la interlengua de aprendientes con lengua materna diferente).

No obstante, en la fase de extensión de las hipótesis, intentaremos conseguir un número de aprendientes superior, de centros educativos diferentes y sobre todo, pondremos en marcha una mejor organización del tiempo que las EOI nos conceden, ya que supone uno de los elementos clave para el éxito de la recogida de datos.

[113] Entendiendo por palabras totales la suma de las palabras de todos los participantes: aprendientes y entrevistadores.

Así, siguiendo con los procesos llevados a cabo, nos hemos detenido en una de las fases más importantes de la compilación: la transcripción, pues de ella depende, en parte, el grado de utilización del corpus en aplicaciones posteriores. Para ello, hemos descrito todas las pautas y convenciones necesarias para el desarrollo de la misma y el programa que se utiliza para llevarlo a cabo (*Transana*).

En nuestro caso, intentamos seguir las pautas ya validadas y utilizadas en distintos corpus orales realizados por el equipo del LLI-UAM, que constituyen así una gran parte de nuestro propio modelo de transcripción. Sin embargo, y como ya hemos mencionado, el tipo de datos que vamos a tratar y el objetivo de nuestra investigación, nos llevan a incorporar nuevas marcas y a tratar distintos fenómenos que no se han contemplado en corpus anteriores de distintas tipologías, como los relacionados con la variedad de lengua oral. Encontramos así una de las dificultades de los corpus actuales: la heterogeneidad de criterios de transcripción, que desembocan, por consiguiente, en una menor probabilidad de reutilización y de comparación de los corpus existentes. Pese al perjuicio que esto pueda suponer, creemos que es mejor abogar por una mayor riqueza de la transcripción, ya que supondrá, a largo plazo, una mejor descripción de la situación comunicativa, que redundará en análisis más adecuados y profundos de la interlengua.

Finalmente, describimos los procesos de alineamiento y conversión a XML, muy necesarios, sobre todo este último para conseguir otro de nuestros objetivos: una mayor difusión gracias a la estandarización y a las posibilidades de reutilización en aplicaciones y otras herramientas informáticas. Para ello, contaremos con un conjunto de ficheros de audio en formato WAV y sus correspondientes transcripciones en formato de texto plano (UTF-8) y XML, organizadas en carpetas para cada uno de los niveles expuestos y representados del MCER.

3. ANÁLISIS DE ERRORES EN APRENDIENTES DE FRANCÉS COMO LENGUA EXTRANJERA (FLE) A PARTIR DEL CORPUS CORAF

1. Introducción

Los corpus de aprendientes, como ya se sabe, ayudan a saber más sobre los procesos mentales implicados en la adquisición de una segunda lengua o lengua extranjera, proporcionando una base de datos suficiente para describir las características de la interlengua en sus diferentes fases de desarrollo.

En el presente estudio, nos proponemos conocer mejor la interlengua de los aprendientes de francés L2, de manera a estar en el futuro en condiciones de introducir mejoras en las estrategias de enseñanza y aprendizaje. Nuestra apuesta está en línea con la de Pienemann (Lightbown, 2003: 6), quien tras conocer el orden secuencial habitual de adquisición de una lengua, lanzó la hipótesis de que se debe fomentar el desarrollo de competencias en los aprendientes cuando estos están preparados para asimilarlas, basándose en su observación de que la instrucción es más efectiva cuando se basa en el nivel de interlengua inmediatamente superior al que poseen los aprendientes.

Uno de los caminos que podemos escoger para analizar la interlengua es realizar un análisis de los errores más frecuentes. Siguiendo la metodología de Tono (Frankenberg et al., 2011), realizaremos un Análisis de la Situación Presente o *Present Situation Analysis* (PSA), con el fin de identificar la competencia actual de los aprendientes, así como las diferencias o la distancia existente entre las metas a conseguir en L2 y su grado o nivel de lengua meta real. En un primer momento, nos detendremos en la identificación de las dificultades, por ser uno de los elementos más interesantes de analizar. En investigaciones futuras, completaremos este análisis con otros más acordes con las nuevas tendencias de la lingüística, es decir realizando análisis del discurso, y observando fenómenos de fluidez, y la dimensión pragmáticos que se refiere al uso de estrategias de comunicación.

Siguiendo la metodología de Corder comentada en capítulos anteriores[114], una vez compilado el corpus objeto del análisis, hemos procedido a la identificación de los errores, con vista a ofrecer la descripción y valoración de los mismos. Así, este AE presenta una perspectiva de cada error desde tres planos distintos, lo que, como podrá comprobarse, aporta mucha información sobre el mismo, y caracteriza de forma muy completa las muestras de interlengua. Los tres criterios de observación utilizados son el lingüístico, el descriptivo y el etiológico

Como ya queda indicado, CORAF es un corpus de aproximadamente 61.000 palabras, de las que 33.915 corresponden a los aprendientes. Es, reiteramos, una muestra relativamente pequeña para pretender realizar cualquier tipo de generalización o extrapolación al conjunto de los aprendientes de FLE en contexto educativo, en entorno no-nativo y con lengua materna española. Nos limitaremos, pues, a mostrar en nuestros resultados las tendencias observadas las cuales, como decimos, requieren ser validadas por investigaciones posteriores hechas con una muestra de mayor tamaño.

En este capítulo, comentaremos cómo se ha ido realizando el análisis de errores, para después ir desgranando los primeros resultados obtenidos. Veremos las tendencias de errores en general, por criterios descritos y por niveles, de forma que podamos obtener una panorámica de la situación actual de los aprendientes entrevistados. Finalmente, nos detendremos en las listas de errores frecuentes, combinando los tres criterios analizados y las partes de la oración o categoría gramatical que se ven más afectadas por el error.

2. Objetivos y dificultades del análisis de errores

El análisis de errores que nos planteamos realizar a partir del corpus CORAF no pretende ser un estudio concluyente sobre las dificultades de los aprendientes hispanohablantes de FLE en general. Su objetivo fundamental es el de realizar un diagnóstico, que nos permita

[114] Dicha metodología prevé el desarrollo del AE en seis fases: recopilación de muestras de lengua, identificación, descripción, explicación, evaluación de los errores y en casi de AE con una orientación didáctica, proposición de actuaciones para la resolución de los errores.

comprender mejor las dificultades del público meta, ante los desafíos del proceso de adquisición de la lengua (Porquier, 1977). Es también lo que hemos denominado anteriormente como análisis de la situación presente (PSA).

Nuestro objetivo es, por tanto, conseguir mostrar las tendencias más sobresalientes que subyacen en el discurso de nuestros aprendientes, de forma que podamos generar hipótesis susceptibles de ser validadas en futuros proyectos e investigaciones.

La metodología de realización del AE incluye identificar, categorizar o describir, explicar y evaluar los errores. El trabajo principal consiste en identificar todos aquellos errores de competencia que se presenten en el transcurso de las entrevistas. No atendemos, como ya hemos mencionado anteriormente, a aquellos fenómenos de la lengua oral que puedan alterar la estructura sintáctica, ya que, como interacción de registro informal que es, los consideramos adecuados, siendo incluso interesantes para la evaluación final de la competencia del aprendiente.

Sin embargo, en ocasiones, identificar un error no tan fácil como se podría suponer, bien porque puede ser descrito de diversas maneras, bien porque al tratarse de una producción oral, la pronunciación puede generar una confusión en cuanto a su interpretación. En ambos casos, no sólo nos encontramos ante errores que calificamos como *ambiguos*, sino que es posible, además, que no sepamos distinguir un error de competencia del que no lo es. En otras palabras, podemos señalar que existen algunos errores en los que no tenemos la certeza de distinguir un problema de competencia lingüística del aprendiente, puesto que la pronunciación incorrecta o incompleta del enunciado nos hace dudar de si se trata de un error que afecta al conocimiento de la lengua, o si por el contrario, se trata de una desviación producida por la situación comunicativa. En algunos casos, el resto del enunciado (o el contexto) puede ayudarnos a su identificación, pero no siempre podemos contar con este factor para las tareas de desambiguación.

Además, es también posible que muchos de estos errores sólo aparezcan en la producción oral del aprendiente, y que no se den en sus producciones escritas. Las características propias de la interacción oral, y entre ellas, la inmediatez y espontaneidad, juegan en contra del aprendiente, que si no tiene aún una habilidad de expresión oral muy desarrollada, al no tener tiempo para reflexionar, tiende a simplificar los

enunciados y, sobre todo, a incurrir en errores que no cometería en otros contextos o tipos de discurso.

En nuestro corpus hemos detectado algunas ambigüedades recurrentes en los siguientes casos:

- Entre el artículo definido masculino singular '*le*' y la forma plural, '*les'*. Lo mismo ocurre entre '*de'* (preposición) y el plural '*des*'.
- Entre formas conjugadas del pasado imperfecto '*j'étais*' y pasado compuesto '*j'ai été'*.
- Entre el pronombre sujeto '*je*' y la forma conjugada del verbo '*avoir*' para la primera persona del singular en presente, '*j'ai*'.
- Entre el artículo indefinido en sus formas femenina '*une*' y masculina '*un*'[115].
- Entre formas del adjetivo de género masculino y femenino, por ejemplo, '*grand*' y '*grande*', o '*intelligent*' y '*intelligente*'[116].

Es evidente que muchos de ellos contemplan problemas de distinción entre fonemas como /ə/ y /ɛ/. Numerosos casos pueden ser por tanto descritos como errores de fonología, ya que reflejan una producción deficiente, pero al tratarse de elementos con importantes distinciones fonológicas, pueden hacernos dudar de si detrás de este primer error, existen otros ligados a aspectos gramaticales. Por lo que, siguiendo la taxonomía expuesta, en realidad, no se trataría sólo de un error fonético, sino también gramatical.

En todos estos casos, acudimos primero al contexto inmediato del enunciado, y posteriormente, al resto de categorías de la oración con los que podamos comparar la producción. Si el error es repetitivo, o si el contexto posee incoherencias, evidentemente, podría tratarse de un error

[115] En este caso nos encontramos con un error fonético claro, pero en muchos de los casos, no sabemos si además podría existir un problema de concordancia de género con el sustantivo al que acompañan.

[116] Para este caso ocurre lo mismo que para el anterior. No podemos afirmar que no oculte también un problema de concordancia de género.

gramatical. Si no, podríamos concluir que es de orden fonológico, o bien, si no es posible encontrar indicios, se señalará como ambiguo.

Finalmente, cabe destacar que la categorización del origen o causa del error (criterio etiológico) implica, en muchos casos, una interpretación bastante subjetiva por parte del investigador. Excepto para los casos de interferencia de la lengua materna o de una tercera lengua (errores interlingüísticos), donde la forma nos permite ver más claramente su origen, el resto de categorías intralingüísticas, supone un importante ejercicio de interpretación de los procesos internos del aprendiente. Además, la distinción entre numerosas causas tampoco facilita la discriminación, ya que existen muchos procesos que pueden interrelacionarse o que pueden afectar al mismo error. Por tanto, el criterio etiológico es el aspecto del análisis que más problemas plantea en nuestro estudio, y que, además, nos dota de unos resultados demasiado específicos.

3. Primeros resultados generales

En el desarrollo de nuestro análisis de errores, hemos tratado de clasificar todos los errores que aparecían en los enunciados de los aprendientes de las 30 entrevistas recogidas en CORAF. Así, en las 33.915 palabras de aprendientes que componen nuestro corpus, hemos de destacar que se detectan aproximadamente unos 1.400 errores de competencia[117].

Valorando las muestras de lengua de los aprendientes de nuestro corpus, podemos subrayar que se caracteriza en términos generales por:

- Un número de errores más elevado de tipo gramatical, seguidos de aquellos relativos al léxico y la sintaxis.
- La parte de la oración más afectada en el conjunto, sin relacionarla con todos los tipos y subtipos de errores analizados, es el verbo, seguido de los artículos, las preposiciones, y el sustantivo, entre otros. La menos recurrente es la conjunción, porque el discurso adolece de

[117] Volvemos a mencionar aquí que se han tenido en cuenta los errores que no están asociados (o que parecen no estarlo) a la situación de producción, obviando así los errores de actuación/producción y otras alteraciones de la misma.

subordinadas complejas de diversos tipos, basándose en las más sencillas y fácilmente adquiridas por los aprendientes (como las de coordinación o las causales con la conjunción *"parce que"*).

- Los errores desde el punto de vista descriptivo se producen, en primer lugar, por una forma errónea, seguidos de la falsa selección, la omisión, la adición y finalmente, la falsa colocación u orden alterado.
- La causa principal de los errores es de orden intralingüístico, aunque con poca distancia con respecto a los de tipo interlingüístico. Existe también un número elevado de errores de causa ambigua o desconocida.
- Aparición de un número desigual de errores, que va decreciendo y aumentando sin una tendencia clara. Pero se comprueba que a mayor número de palabras, y por tanto, en aprendientes de nivel más avanzado, su frecuencia de aparición es menor.

Así, podemos repartir el total de los errores detectados en los distintos niveles representados en el corpus de la forma siguiente:

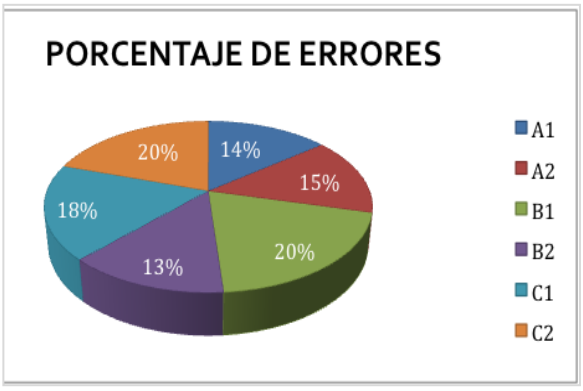

Gráfico 12: Distribución en porcentajes por niveles del total de errores.

Partiendo de los datos obtenidos, el nivel del MCER donde observamos un menor número de errores detectados es el B2, que con sólo 189 de ellos, parece ser el más eficaz en la interacción oral. Resulta curioso que el nivel que contiene más formas desviadas sea el C2 (bien es cierto que sólo dos errores lo separan del nivel B1, el segundo con más errores). De todas formas, ocurre si lo comparamos sólo sobre el total de errores detectados, ya que, evidentemente, si lo relacionamos con el número de palabras producidas por cada nivel (número o valor normalizado), el resultado es distinto, obteniendo porcentajes menores y más adecuados a la realidad. Teniendo en cuenta el valor normalizado se puede apreciar claramente que la proporción de errores por palabras producidas va disminuyendo progresivamente, aunque experimenta un leve repunte en el nivel C2. Este aumento se debe, sin duda, a las características y peculiaridades de los aprendientes entrevistados.

NIVEL	Nº ERRORES	Nº PALABRAS	Nº NORMALIZADO
A1	203	2.506	12,22
A2	207	4.110	19,85
B1	279	4.908	17,59
B2	189	6.858	36,28
C1	254	7.868	30,97
C2	281	7.666	27,28

Tabla 11: Resumen de errores detectados y palabras producidas por niveles del MCER.

Si atendemos a los datos del valor normalizado, cuanto mayor es el número, menor es la distribución relativa de los errores. Es decir, aquellos niveles con un número mayor son aquellos que menos errores cometen, ya que estos aparecen con menor frecuencia en el discurso. Así, el nivel B2 sería el que menos errores produce en la actividad de

interacción oral (cada 36,28 palabras), seguido del nivel C1 y del C2. Los peores porcentajes corresponden, como cabe esperar, a los niveles de usuarios principiantes.

3.1 Clasificación de errores según el criterio lingüístico

El primer paso en el análisis de los resultados es determinar qué tipo de errores son los más usuales en nuestro corpus desde el punto de vista lingüístico.

Este criterio nos aporta información sobre en qué plano del discurso recae la mayor parte de los errores, y si, además, lo relacionamos con la parte de la oración o categoría gramatical sobre la que inciden, podemos saber cuál suele ser la más afectada y mediante qué tipo de errores.

Basándonos en los resultados obtenidos en el AE del corpus CORAF, obtenemos que los errores más frecuentes son de tipo gramatical[118]. Así, el reparto de los distintos tipos estudiados desde el punto de vista lingüístico quedaría como sigue:

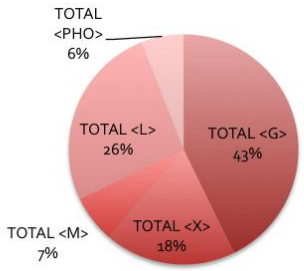

Gráfico 13: Reparto de los errores según análisis por criterio lingüístico

[118] La distinción de las tipologías de errores puede observarse en el capítulo 6, en la página 164, donde comentamos las taxonomías que se utilizan para cada uno de los criterios. En el caso del criterio lingüístico, aplicamos una adaptación de la taxonomía utilizada por Granger (2003a) para el análisis automático de errores del corpus FRIDA, que realiza esta separación entre errores gramaticales, léxicos, sintácticos y morfológicos.

Como podemos observar, los errores más frecuentes son los gramaticales (<G>), seguidos de los relativos al léxico (<L>), a la sintaxis (<X>) y a la morfología (<M>). Los menos frecuentes son los que suponen alteraciones en aspectos fonéticos[119] (<PHO>).

Ciertamente, los errores se reparten de forma desigual a lo largo de los distintos niveles que componen nuestro corpus. En la siguiente figura, podemos observar su distribución normalizada en los diferentes niveles contenidos en CORAF:

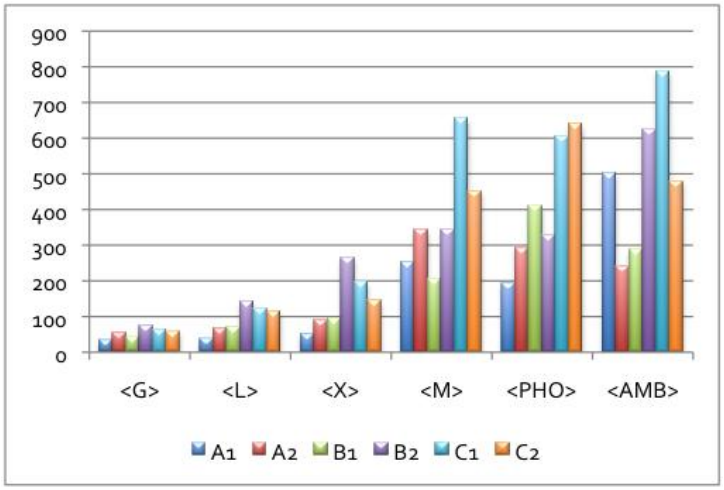

Gráfico 14: Distribución normalizada de tipos de error lingüístico por niveles del MCER.

Siguiendo lo expuesto en el gráfico anterior, los errores, sean del tipo que sean, están presentes en cada uno de los niveles, si bien su presencia es mayor o menor en función del tipo que se analice. En la distribución normalizada que presentamos, aquellos niveles con menor distribución relativa son los que más errores cometen (por extensión, aquellos con números más altos son los que más adecuadamente

[119] Hemos de señalar que en nuestro primer estudio atendemos primero a errores más estructurales, señalando como errores fonéticos aquellos que reflejan desviaciones especialmente significativas.

producen su discurso, porque los errores aparecen de forma menos frecuente).

El nivel A1 es lógicamente el que más errores de todos los tipos presenta, exceptuando los errores ambiguos, donde el nivel A2 es el que produce más formas desviadas.

Los niveles más regulares son, por tanto, B2, C1 y C2, que corresponden a aquellos de competencia más avanzada. De todos ellos, el nivel B2 es el que mejores resultados ofrece en los de tipo gramatical, sintáctico y léxico. Resulta curioso que los aprendientes de este nivel cometan menos errores que los más avanzados. La explicación a esta tendencia tiene su origen en las características de los hablantes que forman parte de nuestras muestras. En el nivel C2 tenemos algunos hablantes que producen su discurso de forma fluida pero sin atender a la corrección. Sin embargo, en el nivel B2 existen dos aprendientes que producen de forma muy correcta, en parte, porque además de recibir la instrucción adecuada a su nivel en entorno educativo, han podido interactuar de forma prolongada en países francófonos (por medio de programas de movilidad *Erasmus* o en estancias en programas de verano). Sin duda, sería necesario ampliar las muestras de hablantes para poder determinar si esta tendencia se mantiene, o si, por el contrario, como cabría de esperar, los niveles más avanzados son los que menos errores deberían de cometer en todos los tipos.

La gráfica es más acorde con las tendencias esperadas en los errores de tipo morfológico, fonético y ambiguo, donde los niveles C1 y C2 son los que menos errores cometen, superando al resto de niveles. Así, de ellos dos, el nivel C1 es el más competente, aunque, como hemos comentado anteriormente, es posible que se deba a las características de los hablantes de C2, que no han adquirido aún el nivel que se les presupone (como hemos comentado ya, el momento en el que se graban nuestras entrevistas se sitúa antes de acabar el curso académico).

En un análisis de errores de tipo lingüístico, también podemos observar la(s) categoría(s) gramatical(es) que resulta más afectada(s) por los distintos errores objeto del análisis. Así, y como en la mayoría de AE de aprendientes del francés, en CORAF encontramos las mayores dificultades en los verbos, los artículos, las preposiciones y los sustantivos. Con un menor número de apariciones seguirían los pronombres, los adjetivos, los adverbios, los determinantes y por último, las conjunciones. Es obvio que la lengua utilizada no es demasiado

compleja, por lo que las conjunciones, al no haber muchas oraciones subordinadas, o al utilizar siempre aquellas más conocidas por el aprendiente, apenas sí aparecen en nuestro corpus (y por ende, tampoco presentan graves problemas o un número elevado de errores).

La distribución de las partes de la oración afectadas por errores sería la siguiente[120]:

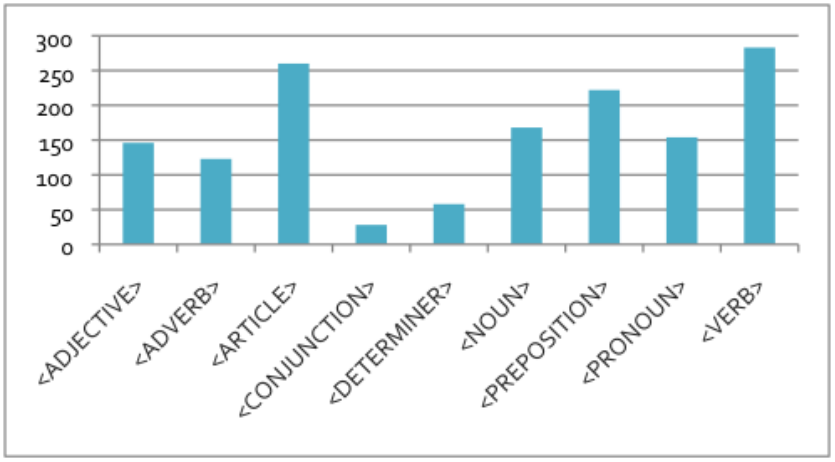

Gráfico 15: Distribución de categorías gramaticales más afectadas por errores presentes en CORAF.

Entre los errores más frecuentes destacamos, después de los verbos, a los artículos. Puede parecer extraño que esta categoría, considerada como de las de adquisición más temprana aparezca como una de las más afectadas. Existen, posiblemente, distintas causas para este fenómeno. Una de ellas es que dicha categoría engloba tanto a artículos definidos como indefinidos, y a otros de más compleja adquisición, como son los partitivos y los contractos (*du, au, des*…). Por un lado, el número de errores aumenta por las deficiencias en la utilización de estos últimos, de mayor dificultad para el aprendiente, y por otro, porque en los artículos definidos e indefinidos están presentes

[120] Dentro de cada una de las categorías propuestas se incluyen las subcategorías presentes en cada uno. Así, por ejemplo, los pronombres abarcan tanto a aquellos personales, como posesivos, relativos, etcétera. Los verbos engloban tanto a los simples como compuestos, así como a las formas en infinitivo o gerundio.

muchos de los errores ambiguos de nuestro corpus. Sobre todo, estos problemas de ambigüedad, como hemos señalado anteriormente, aparecen en la distinción del singular/plural entre los definidos *le/les* y en los artículos indefinidos *un/une*. Toda esta suma de dificultades provoca que, en conjunto, sea una de las categorías más presente en los distintos errores analizados y en cada uno de los niveles.

No obstante, podemos afirmar que nuestro corpus sigue, en líneas generales, la tónica habitual para este tipo de análisis con aprendientes. Así, podemos señalar a Porquier (1977: 27), quien señala, aunque de forma bastante general, que los errores más frecuentes de los aprendientes de FLE, independientemente de cuál sea su lengua materna, son aquellos que tienen que ver con los determinantes, las formas verbales, la morfología del género y número y las preposiciones.

Al igual que en la distribución general antes expuesta, creemos conveniente establecer el reparto de estos errores por niveles del MCER, de forma que podamos apreciar el volumen de estos por nivel, o su desarrollo a lo largo del proceso de adquisición (si bien la precisión de los datos no será la misma que al analizar un corpus de tipo longitudinal). La distribución podría reflejarse de la manera siguiente:

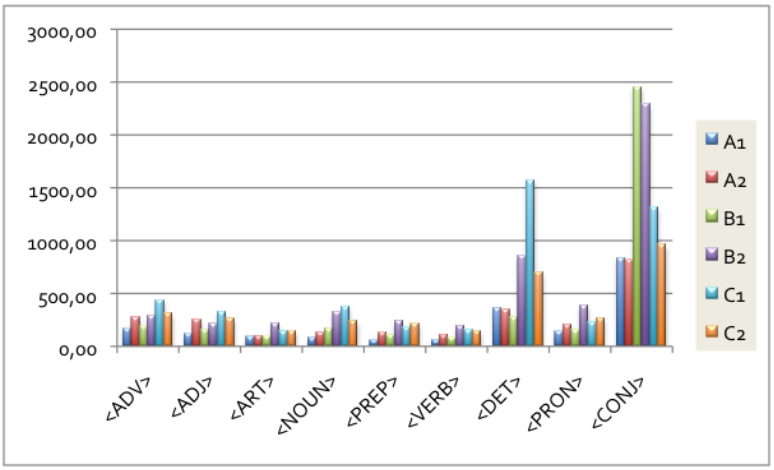

Gráfico 16: Distribución de los errores por categorías gramaticales afectadas en cada uno de los niveles de CORAF.

Como era de esperar, los niveles más avanzados son los que presentan un menor número de errores en las distintas categorías, excepto en el caso de las conjunciones, ya que, evidentemente, el resto de niveles no utilizan oraciones tan complejas, y por lo tanto, no cometen los errores típicos de ese tipo de enunciados. Los niveles que muestran menores dificultades siguen siendo los relativos a B2 y C1.

Así, el **nivel A1** posee más dificultades asociadas a las categorías sustantivo, verbo y artículo. En el caso de los verbos, la mayoría de los problemas se presentan en formas finitas simples y en formas compuestas. Existen también algunas dificultades en formas de uso del infinitivo.

Los artículos, como para el resto de niveles, también suponen un cierto obstáculo. En este caso los mayores problemas se reflejan en los artículos definidos, seguidos de los indefinidos, los contractos y los partitivos.

La distribución de las categorías más afectadas puede observarse de forma más detallada en la siguiente tabla:

CATEGORIA AFECTADA	SUBCATEGORÍA	NIVEL	Nº TOTAL ERRORES	VALOR RELATIVO (% errores/ nivel)	VALOR NORMALIZADO (error /nº palabras totales nivel)
VERBO	FORMA FINITA SIMPLE	A1	25	12,31 %	100,24
	FORMA COMPUESTA		11	5,42 %	227,8
	INFINITIVO SIMPLE		5	2,46 %	501,2
	TOTAL	**A1**	**41**	**20,19 %**	**61,12**
SUSTANTIVO	SIMPLE COMÚN	A1	32	15,76 %	78,31
	TOTAL	**A1**	**32**	**15,76 %**	**78,31**
PREPOSICIÓN	PREPOSICIÓN SIMPLE	A1	36	17,74 %	69,61

	PREPOSICIÓN COMPUESTA		3	1,48 %	835,3
	TOTAL	**A1**	**39**	**19,21 %**	**64,25**
ARTÍCULO	DEFINIDOS	A1	15	7,38 %	167,06
	INDEFINIDOS		7	3,45 %	358
	CONTRACTOS		3	1,48 %	835,3
	PARTITIVOS		3	1,48 %	835,3
	TOTAL	**A1**	**28**	**13,79 %**	**89,5**
ADJETIVO	ADJETIVO SIMPLE	A1	18	8,87 %	139,22
	ADJETIVO COMPARATIVO		3	1,48 %	835,3
	TOTAL	**A1**	**21**	**10,34 %**	**119,3**
DETERMINANTES	INDEFINIDOS	A1	5	2,46 %	501,2
	NUMERALES		2	0,98 %	1253
	TOTAL	**A1**	**7**	**3,45 %**	**358**
PRONOMBRE	PERSONAL	A1	16	7,88 %	156,6
	INDEFINIDO		1	0,05 %	2.506
	DEMOSTRATIVO		1	0,05 %	2.506
	TOTAL	**A1**	**18**	**8,87 %**	**139,22**
ADVERBIO	SIMPLE	A1	15	7,38 %	167,1
	TOTAL	**A1**	**15**	**7,38 %**	**167,1**
CONJUNCIÓN	SIMPLE	A1	3	1,48 %	835,3
	TOTAL	**A1**	**3**	**1,48 %**	**835,3**

Tabla 12: Valores y porcentajes de errores de las categorías gramaticales para el nivel A1.

El **nivel A2**, usuario básico plataforma según el MCER, posee sus mayores dificultades en el uso de los determinantes. De ellos, destacan sus problemas con los determinantes numerales, indefinidos, demostrativos y posesivos.

Los artículos también suponen una categoría bastante afectada por sus errores, con prácticamente igual número para los indefinidos y los definidos, pero sin embargo, con mayor presencia de dificultades que el nivel A1 en partitivos y contractos.

El resto de categorías se reparten de la siguiente forma:

CATEGORIA AFECTADA	SUBCATEGORÍA	NIVEL	Nº TOTAL ERRORES	VALOR RELATIVO (% errores/ nivel)	VALOR NORMALIZADO (error /nº palabras totales nivel)
SUSTANTIVO	SIMPLE COMÚN	A2	28	13,53 %	146,78
	PROPIO		1	0,04 %	4.110
	COMPUESTO		1	0,04%	4.110
	TOTAL	**A2**	**30**	**14'49 %**	**137**
VERBO	FORMA FINITA SIMPLE	A2	21	10,14 %	195,71
	FORMA COMPUESTA		9	4,35 %	456,6
	INFINITIVO SIMPLE		5	2,41 %	822
	INFINITIVO COMPUESTO		2	0,96 %	2.055
	TOTAL	**A2**	**37**	**17,87 %**	**111,08**
PREPOSICIÓN	PREPOSICIÓN SIMPLE	A2	28	13,53 %	146,78
	PREPOSICIÓN COMPUESTA		3	1,45 %	1.370
	TOTAL	**A2**	**31**	**15 %**	**132,58**
ARTÍCULO	DEFINIDOS	A2	10	4,83 %	411
	INDEFINIDOS		11	5,31 %	373,63
	CONTRACTOS		10	4,83 %	411
	PARTITIVOS		9	4,35 %	456,6
	TOTAL	**A2**	**41**	**19,8 %**	**100,24**
ADJETIVO	ADJETIVO SIMPLE	A2	15	7,24 %	274
	TOTAL	**A2**	**15**	**7,24 %**	**274**
DETERMINANTES	INDEFINIDOS	A2	3	1,45 %	1.370
	NUMERALES		4	1,93 %	1.027,5
	DEMOSTRATIVO		2	0,96 %	2.055
	POSESIVO		2	0,96 %	2.055
	TOTAL	**A2**	**11**	**5,31 %**	**373,63**
PRONOMBRE	PERSONAL	A2	13	6,28 %	316,15
	DEMOSTRATIVO		2	0,96 %	2.055
	EXCLAM.-INTERR.		2	0,96 %	2.055
	RELATIVO		2	0,96 %	2.055
	INDEFINIDO		1	0,04%	4.110

	TOTAL	**A2**	**20**	**9,6 %**	**205,5**
ADVERBIO	SIMPLE	A2	13	13	6,28 %
	COMPUESTO		2	0,96 %	2.055
	TOTAL	**A2**	**15**	**7,24 %**	**274**
CONJUNCIÓN	SUB. SIMPLE	A2	4	1,93 %	1.027,5
	COORDINACIÓN		1	0,04%	4.110
	TOTAL	**A2**	**5**	**2,41 %**	**822**

Tabla 13: Valores y porcentajes de errores de las categorías gramaticales para el nivel A2.

El **nivel B1** observamos un mayor número de dificultades en la casi totalidad de las categorías analizadas, destacando, sobre todo, en verbos y preposiciones, donde es el nivel que más errores aporta al conjunto.

En primer lugar, entre sus problemas en el uso y formación de verbos, destacamos un mayor número de errores en las formas finitas simples, seguido de los de formas compuestas y una menor presencia en las formas de infinitivo.

En cuanto a las preposiciones, los principales problemas surgen en preposiciones simples, y sólo 4 errores en las compuestas (de hecho, es poco habitual el uso de preposiciones de dicho tipo en todos los niveles).

Aunque menos significativos para el conjunto de niveles, los aprendientes de B1 también poseen dificultades en los adjetivos, donde además de problemas en las formas simples, reflejan dificultades en la formación de comparativos (relacionadas, en este caso, con algunos de tipo irregular, como *mieux*, *meilleur*, y *pire*).

Además, siguen encontrando ciertos obstáculos en otras categorías como los determinantes (destacando el uso de los indefinidos, los posesivos y los demostrativos) y los pronombres. En estos últimos las mayores dificultades recaen sobre los pronombres personales, que destacan sobre el resto de subtipos: relativos, indefinidos, exclamativos-interrogativos, impersonales y demostrativos.

Podemos observar todos los datos referentes al nivel B1 en la siguiente tabla:

CATEGORIA AFECTADA	SUBCATEGORÍA	NIVEL	Nº TOTAL ERRORES	VALOR RELATIVO (% errores/ nivel)	VALOR NORMALIZADO (error /nº palabras totales nivel)
SUSTANTIVO	SIMPLE COMÚN	B1	26	9,31 %	188,7
	PROPIO		3	1,07 %	1.636
	TOTAL	**B1**	**29**	**10,3 %**	**169,2**
VERBO	FORMA FINITA SIMPLE	B1	35	12,5 %	140,22
	FORMA COMPUESTA		15	5,3 %	327,2
	INFINITIVO SIMPLE		5	1,79 %	981,6
	INFINITIVO COMPUESTO		1	0,3 %	4.908
	GERUNDIO		1	0,3 %	4.908
	TOTAL	**B1**	**57**	**20,43 %**	**86,10**
PREPOSICIÓN	PREPOSICIÓN SIMPLE	B1	40	14,3 %	122,7
	PREPOSICIÓN COMPUESTA		4	1,43 %	1.227
	TOTAL	**B1**	**44**	**15,7 %**	**111,5**
ARTÍCULO	DEFINIDO	B1	17	6,09 %	288,7
	INDEFINIDO		19	6,81 %	258,3
	CONTRACTO		7	2,5 %	701,1
	PARTITIVO		7	2,5 %	701,1
	TOTAL	**B1**	**50**	**17,9 %**	**98,16**
ADJETIVO	ADJETIVO SIMPLE	B1	27	9,6 %	181,77
	ADJETIVO COMPARATIVO		4	1,43 %	1.227
	TOTAL	**B1**	**31**	**11,1 %**	**158,3**
DETERMINANTE	INDEFINIDO	B1	5	1,79 %	981,6
	NUMERAL		3	1,07 %	1.636
	POSESIVO		4	1,43 %	1.227
	DEMOSTRATIVO		3	1,07 %	1.636
	TOTAL	**B1**	**15**	**5,3 %**	**327,2**
PRONOMBRE	PERSONAL	B1	21	7,53 %	233,7
	RELATIVO		5	1,79 %	981,6
	INDEFINIDO		2	0,7 %	2.454

	EXCL.-INTERR.		1	0,3 %	4.908
	DEMOSTRATIVO		1	0,3 %	4.908
	IMPERSONAL		1	0,3 %	4.908
	TOTAL	**B1**	**31**	**11,1 %**	**158,3**
ADVERBIO	SIMPLE	B1	26	9, 31 %	188,7
	TOTAL	**B1**	**26**	**9,31 %**	**188,7**
CONJUNCIÓN	SUB. SIMPLE	B1	1	0,3 %	4.908
	TOTAL	**B1**	**1**	**0,3 %**	**4.908**

Tabla 14: Valores y porcentajes de errores de las categorías gramaticales para el nivel B1.

El **nivel B2,** usuario independiente avanzado para el MCER, es el nivel más regular, puesto que posee dificultades en todas las categorías, como ocurre con el resto, pero no en un número importante o significativo. Quizá la única categoría donde destaquen más dificultades sea en los adjetivos, y generalmente, debido a problemas con la concordancia del género y número y en la formación errónea o un uso inadecuado (sobre todo en fenómenos de *falsos amigos,* con formas típicas como *semblable/pareil* o *rare/bizarre* y de creación léxica, donde surgen problemas por una interferencia de la L1 al realizar procesos de sufijación o prefijación –**traditionnal/traditionnel*-). Podemos ver los errores más frecuentes por categorías gramaticales en la siguiente tabla:

CATEGORIA AFECTADA	SUBCATEGORÍA	NIVEL	Nº TOTAL ERRORES	VALOR RELATIVO (% errores/ nivel)	VALOR NORMALIZADO (error /nº palabras totales nivel)
SUSTANTIVO	SIMPLE COMÚN	B2	19	10,05 %	360,95
	COMPUESTO		2	1,06 %	3429
	TOTAL	**B2**	**21**	**11,1 %**	**326,5**
VERBO	FORMA FINITA SIMPLE	B2	16	8,4 %	428,62
	FORMA COMPUESTA		17	8,9 %	403,41

	INFINITIVO SIMPLE		3	1,5 %	2.286
	TOTAL	**B2**	**36**	**19,05 %**	**190,5**
PREPOSICIÓN	PREPOSICIÓN SIMPLE	B2	26	13,7 %	263,76
	PREPOSICIÓN COMPUESTA		3	1,5 %	2.286
	TOTAL	**B2**	**29**	**15,34 %**	**236,4**
ARTÍCULO	DEFINIDO	B2	11	5,8 %	623,4
	INDEFINIDO		18	9,5 %	381
	PARTITIVO		3	1,5 %	2.286
	TOTAL	**B2**	**32**	**16,93 %**	**214,3**
ADJETIVO	ADJETIVO SIMPLE	B2	22	11,64 %	311,72
	ADJETIVO COMPARATIVO		2	1,06 %	3.429
	TOTAL	**B2**	**24**	**12,7 %**	**285,7**
DETERMINANTE	INDEFINIDO	B2	1	0,5 %	6.858
	NUMERAL		5	2,6 %	1.371,6
	POSESIVO		2	1,06 %	3429
	TOTAL	**B2**	**8**	**4,2 %**	**857,2**
PRONOMBRE	PERSONAL	B2	12	6,35 %	571,5
	RELATIVO		3	1,5 %	2.286
	DEMOSTRATIVO		2	1,06 %	3.429
	IMPERSONAL		1	0,5 %	6.858
	TOTAL	**B2**	**18**	**9,5 %**	**381**
ADVERBIO	SIMPLE	B2	17	8,9 %	403,41
	TOTAL	**B2**	**17**	**8,9 %**	**403,41**
CONJUNCIÓN	SUB. SIMPLE	B2	3	1,5 %	2.286
	TOTAL	**B2**	**3**	**1,5 %**	**2.286**

Tabla 15: Valores y porcentajes de errores de las categorías gramaticales para el nivel B2.

El **nivel C1** también posee varias categorías en las que podemos destacar distintos problemas. Es el caso de los artículos, las preposiciones, los pronombres y las conjunciones, que aparecen por primera vez entre las dificultades más importantes. Evidentemente, el

discurso va siendo cada vez más complejo, y por lo tanto, las conjunciones empiezan a tener más presencia en él.

Así, en cuanto a los artículos, las mayores dificultades residen en los indefinidos, aunque con una distancia pequeña con respecto a los definidos. También se encuentran algunos problemas con artículos contractos, pero disminuyen considerablemente aquellos relacionados con los partitivos (apareciendo sólo una vez).

En lo referente a las preposiciones, como en el resto de niveles, los problemas se sitúan en las de tipo simple, ya que las formas compuestas son bastante menos frecuentes (existen sólo 4 formas erróneas, un 1,57 % del total de errores del nivel C1).

Por otra parte, cabe destacar las dificultades que presentan los pronombres. La mayoría de ellos tiene lugar en los de tipo personal, y se relacionan en gran número con la omisión en el sujeto de la oración, y en pronombres necesarios en posición preverbal. El resto de dificultades aparecen en los pronombres relativos y en los demostrativos.

Todos estos datos aparecen reflejados en la siguiente tabla:

CATEGORIA AFECTADA	SUBCATEGORÍA	NIVEL	Nº TOTAL ERRORES	VALOR RELATIVO (% errores/ nivel)	VALOR NORMALIZADO (error /nº palabras totales nivel)
SUSTANTIVO	SIMPLE COMÚN	C1	18	7,08 %	437,1
	PROPIO		3	1,18 %	2.622,6
	TOTAL	**C1**	**21**	**8,26 %**	**374,6**
VERBO	FORMA FINITA SIMPLE	C1	31	12,2 %	253,8
	FORMA COMPUESTA		14	5,5 %	562
	INFINITIVO SIMPLE		3	1,18 %	2.622,6
	INFINITIVO COMPUESTO		1	0,3 %	7868
	TOTAL	**C1**	**50**	**19,6 %**	**157,3**
PREPOSICIÓN	PREPOSICIÓN SIMPLE	C1	39	15,3 %	201,7

	PREPOSICIÓN COMPUESTA		4	1,57 %	1.967
	TOTAL	**C1**	**43**	**16,9 %**	**182,9**
ARTÍCULO	DEFINIDO	C1	21	8,26 %	374,66
	INDEFINIDO		22	8,66 %	357,6
	CONTRACTO		10	3,9 %	786,8
	PARTITIVO		1	0,3 %	7868
	TOTAL	**C1**	**54**	**21,2 %**	**145,7**
ADJETIVO	ADJETIVO SIMPLE	C1	22	8,66 %	357,6
	ADJETIVO COMPARATIVO		2	0,8 %	3934
	TOTAL	**C1**	**24**	**9,4 %**	**327,8**
DETERMINANTE	INDEFINIDO	C1	2	0,8 %	3934
	POSESIVO		2	0,8 %	3934
	DEMOSTRATIVO		1	0,3 %	7868
	TOTAL	**C1**	**5**	**1,9 %**	**1573,6**
PRONOMBRE	PERSONAL	C1	25	9,8 %	314,72
	RELATIVO		8	3,14 %	983,5
	DEMOSTRATIVO		2	0,8 %	3934
	TOTAL	**C1**	**35**	**13,77**	**224,8**
ADVERBIO	SIMPLE	C1	17	6,6 %	462,82
	COMPUESTO		1	0,3 %	7868
	TOTAL	**C1**	**18**	**7,08 %**	**437,1**
CONJUNCIÓN	SUB. SIMPLE	C1	4	1,57 %	1967
	COORDINACIÓN		2	0,8 %	3934
	TOTAL	**C1**	**6**	**2,36 %**	**1311,3**

Tabla 16: Valores y porcentajes de errores de las categorías gramaticales para el nivel C1.

Por último, el **nivel C2** o usuario competente (maestría), posee un número destacado de dificultades en cuatro grandes categorías: artículos, sustantivos, verbos y conjunciones.

En lo que respecta a los artículos, los indefinidos suponen más problemas a los aprendientes de C2 de CORAF que los definidos, y en menor medida los partitivos y los contractos.

Los sustantivos son quizá una de las categorías más destacadas, junto con los verbos, y en este caso, los grandes problemas residen en la inapropiada selección del léxico al contexto de uso y en la creación léxica, que tiene como resultado distintas formas erróneas.

Por otra parte, los verbos también suponen una categoría problemática, destacando las formas simples sobre las compuestas y las de infinitivo simple.

Vemos también que las conjunciones empiezan a aparecer con mayor frecuencia, como ocurría con el nivel C1, lo que supone, además, mayores dificultades que para el resto de niveles, que apenas si usan esta categoría en su discurso (si bien en el nivel C2 tampoco tienen una presencia muy significativa). En el caso de las conjunciones, los problemas se concentran en el uso de las conjunciones simples y las coordinadas.

Así, podemos ver todos los valores de forma detallada en la siguiente tabla:

CATEGORIA AFECTADA	SUBCATEGORÍA	NIVEL	Nº TOTAL ERRORES	VALOR RELATIVO (% errores/ nivel)	VALOR NORMALIZADO (error /nº palabras totales nivel)
SUSTANTIVO	SIMPLE COMÚN	C2	31	11 %	247,29
	PROPIO		1	0,35 %	7666
	TOTAL	**C2**	**32**	**11,3 %**	**239,56**
VERBO	FORMA FINITA SIMPLE	C2	37	13,16 %	207,18
	FORMA COMPUESTA		12	4,27 %	638,83
	INFINITIVO SIMPLE		6	2,13 %	1277,6
	TOTAL	**C2**	**55**	**19,57 %**	**139,38**
PREPOSICIÓN	PREPOSICIÓN SIMPLE	C2	36	12,8 %	212,9
	TOTAL	**C2**	**36**	**12,8 %**	**212,9**
ARTÍCULO	DEFINIDO	C2	20	7,11 %	383,3
	INDEFINIDO		24	8,54 %	319,4
	CONTRACTO		5	1,78 %	1533,2

	PARTITIVO		4	1,42 %	1916,5
	TOTAL	**C2**	**53**	**18,86 %**	**144,64**
ADJETIVO	ADJETIVO SIMPLE	C2	24	8,54 %	319,4
	ADJETIVO COMPARATIVO		5	1,78 %	1533,2
	TOTAL	**C2**	**29**	**10,35 %**	**264,3**
DETERMINANTE	INDEFINIDO	C2	2	0,7 %	3833
	NUMERAL		3	1,06 %	2555,3
	POSESIVO		3	1,06 %	2555,3
	DEMOSTRATIVO		3	1,06 %	2555,3
	TOTAL	**C2**	**11**	**3,9 %**	**696,9**
PRONOMBRE	PERSONAL	C2	16	5,6 %	479,1
	DEMOSTRATIVO		4	1,42 %	1916,5
	RELATIVO		4	1,42 %	1916,5
	INDEFINIDO		2	0,7 %	3833
	EXCL.-INTERR.		2	0,7 %	3833
	POSESIVO		1	0,35 %	7666
	TOTAL	**C2**	**29**	**10,35 %**	**264,3**
ADVERBIO	SIMPLE	C2	23	8,18 %	333,3
	COMPUESTO		2	0,7 %	3833
	TOTAL	**C2**	**25**	**9,96 %**	**306,6**
CONJUNCIÓN	SUB. SIMPLE	C2	5	1,78 %	1533,2
	COORDINACIÓN		3	1,06 %	2555,3
	TOTAL	**C2**	**8**	**2,8 %**	**958,2**

Tabla 17: Valores y porcentajes de errores de las categorías gramaticales para el nivel C2.

3.2. Clasificación de errores por criterio descriptivo

El criterio descriptivo es otro de los tipos de análisis que establecemos en nuestro estudio. En este caso, dicho criterio sobrepasa ya los límites de lo lingüístico y aporta más información sobre el error, mostrándonos qué desviación se produce su forma en relación a la que

consideraríamos adecuada en la lengua meta. Así, señalamos en nuestra taxonomía cinco grandes procesos: adición, omisión, falsa selección, forma errónea y colocación falsa u orden incorrecto.

Analizados los resultados, en el conjunto de CORAF la gran mayoría de errores se producen por una formación errónea (<MIF>). Es decir, las formas contienen incorrecciones con respecto al enunciado en la lengua meta que podría producir un hablante nativo.

Posteriormente, el resto de errores se reparten entre los de falsa selección (<MIS>) (muy presente en el uso de las preposiciones y otras categorías funcionales), omisión (<OM>) (relacionado con el *olvido* del sujeto y la partícula *'pas'* en la negación), y en menor medida, los de adición (<AD>) y falsa colocación (<WRO>).

La distribución total se puede representar de la siguiente manera:

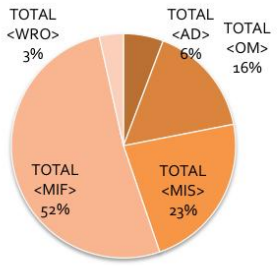

Gráfico 17: Distribución del porcentaje total de errores según el criterio descriptivo.

Por niveles representados en CORAF, la distribución normalizada de los errores que se observa es la siguiente:

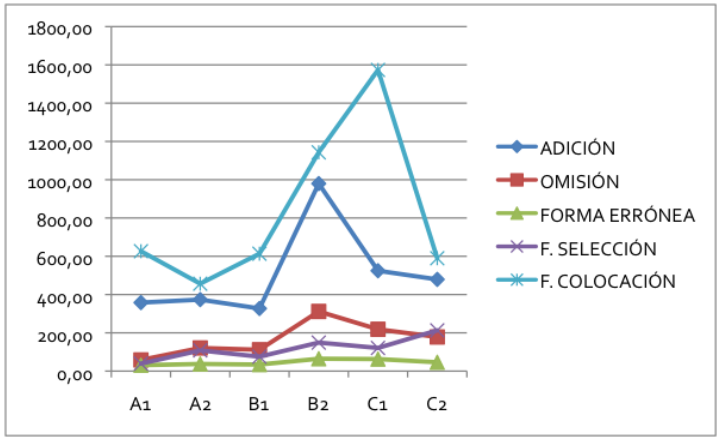

Gráfico 18: Distribución normalizada de los errores en el corpus CORAF según el criterio descriptivo.

Los errores desde el punto de vista descriptivo vemos que se distribuyen de forma clara, destacando por encima de todos ellos los relativos a la forma errónea, que está muy presente y es el más numeroso en todos los niveles. El resto de tipos se agrupa paralelamente, siendo el menos frecuente, como ya hemos indicado, el de falsa colocación u orden incorrecto, seguido del proceso de adición.

Los distintos subtipos no guardan un desarrollo o reparto gradual según los niveles, aumentando y decreciendo de forma bastante irregular. Tan sólo encontramos un decrecimiento en el de falsa selección, que es el único que se reduce al llegar al nivel C2.

En resumen, podemos afirmar que en cuanto al criterio descriptivo, la forma errónea es el proceso más habitual, para todos los niveles y prácticamente con todas las categorías de la oración, aunque más acusado en el uso de verbos (168 apariciones, lo que supone un 12 % de los errores totales y un error cada 201,87 palabras), sustantivos (94 errores, un 6,7 % del total y un error cada 360,8 palabras) y adjetivos (95 apariciones, un 6,7 % del total de errores, apareciendo cada 357 palabras).

Por su parte, los procesos de falsa selección son más frecuentes en categorías funcionales, destacando sobre todas ellas las preposiciones (114). Y también en problemas de orden semántico con sustantivos (46

errores, que componen el 3,28 % del conjunto de errores, apareciendo cada 737,2 palabras del corpus), estando presente en fenómenos como el uso de léxico inadecuado.

Los errores de omisión se relacionan principalmente con los pronombres personales (43 apariciones, un 3,07 % del total de errores, apareciendo cada 788,7 palabras), con los adverbios (35 errores, un 2,5 % de total, apareciendo cada 969 palabras del corpus) y las preposiciones (29 errores, suponiendo un 2,07 % del conjunto de errores, y apareciendo cada 1169,4 palabras).

Los de adición, mucho menos frecuentes, tienen que ver habitualmente con las preposiciones (23 errores, un 1,6 % del total, con una frecuencia de aparición de 1474,5 palabras) y los pronombres personales (11 errores, un 0,7 % del total, apareciendo cada 3083,2 palabras).

Finalmente, los de falsa colocación u orden incorrecto son más numerosos en el uso de los adverbios (24 errores, un 1,71 % del conjunto de errores, con un valor normalizado de 1413,12).

3.3 Clasificación de errores según el criterio etiológico

El criterio etiológico es sin duda el que más información aporta sobre el error, ya que alude a la causa del mismo y se relaciona estrechamente con los procesos que el aprendiente pone en práctica para desarrollar su sistema lingüístico intermedio o interlengua.

Así, como venimos señalando, es el criterio que más problemas presenta a la hora de categorizar los errores, ya que es muy habitual que el investigador desconozca el proceso que se ha llevado a cabo, o que tenga dudas ante qué subtipo predomina. En el criterio etiológico hace falta interpretar el enunciado y descubrir los procesos que subyacen a su utilización, por lo que no podemos obviar que se trata del criterio más subjetivo de todos aquellos que analizamos.

Como sabemos, nuestra taxonomía diferencia entre tres grandes tipos de errores: *intralingüísticos* (<INTRA>) o referentes a problemas por interferencia de reglas y otros procesos conocidos y adquiridos de la lengua meta; *interlingüísticos* (<INTER>) o problemas por la interferencia

de la lengua materna y otras lenguas conocidas por el aprendiente; y *ambiguos* (<UKN>), aquellos en los que la causa es desconocida o ambigua para el investigador.

Así, una vez analizados los errores presentes en CORAF, existe un mayor número de ellos asociados a interferencias intralingüísticas, aunque no con un número muy superior al siguiente tipo más frecuente, los errores interlingüísticos. Por tanto, podríamos señalar que ambas interferencias están prácticamente equilibradas en nuestro corpus. La distribución total de los mismos se puede observar en la siguiente figura:

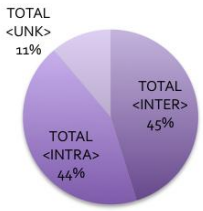

Gráfico 19: Porcentajes de presencia de errores según el criterio etiológico.

En los errores interlingüísticos, la influencia más notable es la de la lengua materna, que abarca el 96'4% de los errores, dejando el porcentaje restante a la influencia de otras lenguas distintas a la materna (el caso más frecuente en CORAF es la interferencia del inglés).

En cuanto a los errores intralingüísticos los errores se subdividen entre los distintos tipos existentes de la siguiente forma:

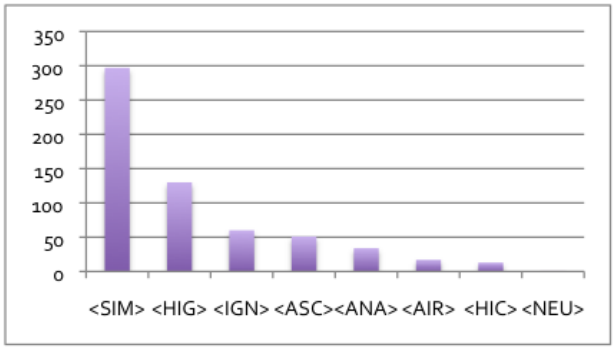

Gráfico 20: Distribución de errores intralingüísticos en el corpus CORAF.

Por tanto, podemos señalar que los procesos intralingüísticos habituales que más interfieren en el discurso oral de los aprendientes de CORAF son los de simplificación (SIM), hipergeneralización (HIG) e ignorancia (IGN) de las restricciones de las reglas. De forma menor, estarían los de asociación cruzada (ASC), analogía (ANA) y aplicación incompleta de las reglas (AIR). Finalmente, en un número poco representativo aparecen los de hipercorrección (HIC, asociado generalmente a verbos compuestos y la formación del *passé composé*) y los de neutralización (<NEU>).

Al igual que para otros criterios, creemos muy útil mostrar cómo se reparten los distintos procesos origen de los errores según los niveles expuestos en CORAF. Así, los resultados obtenidos, pueden observarse en el siguiente gráfico:

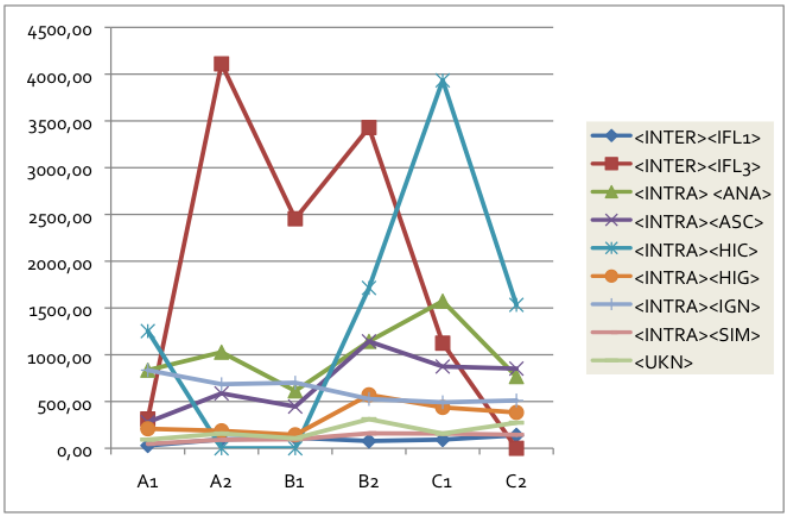

Gráfico 21: Distribución normalizada de los errores según el criterio etiológico en todos los niveles.

Como hemos señalado anteriormente, los errores más frecuentes se producen por la adaptación de procesos intralingüísticos (que

considera la suma de más de siete procesos distintos, como podemos ver en el gráfico adjunto).

Todas las causas posibles de error se manifiestan en mayor o menor medida en todos los niveles del MCER, si bien suelen estar más relacionados con los niveles B1 y C1, que son los que más errores aportan al conjunto. El nivel B2, como ocurre con el resto de criterios, es el nivel más regular, además de ser en el que más errores de tipo intralingüístico confluyen, decayendo los causados por la interferencia de la lengua materna y los ambiguos.

No obstante, si miramos cada uno de los procesos tanto intralingüísticos como interlingüísticos de forma autónoma, los errores más recurrentes se producen por la interferencia de la L1, cuya presencia sigue siendo bastante importante en todos los niveles. Esta interferencia de la L1 se relaciona principalmente con la utilización de sustantivos (97 errores, 6,92% del total de errores, con un valor normalizado de 349,6), preposiciones (67 errores, un 4,78 % del conjunto, y un valor normalizado de 506,1), adverbios (62 errores, conformando un 4,42 % del total, y con un valor normalizado de 547), adjetivos (53 errores, un 3,8 % del conjunto y un valor normalizado de 639,9) y pronombres personales (47 errores, que suponen un 3,3 % del total de errores, con un valor normalizado de 721,5). Es curioso que el nivel C2 siga manteniendo una fuerte presencia de errores de tipo interlingüístico, pero, como ya hemos indicado anteriormente, se debe, ante todo, a las características de los aprendientes que forman parte de este nivel en nuestro corpus, muy tendentes a apoyarse en los conocimientos de la L1 para la construcción de su discurso. De hecho, muchos de los errores de tipo interlingüístico de C2 se relacionan con la creación de léxico y por fenómenos como el del uso de *falsos amigos*.

La interferencia de una tercera lengua es muy reducida, contando con escaso número de casos significativos, y no presente en todos los niveles.

También es importante el número de errores de causa ambigua o desconocida, que es la tercera categoría con más presencia en el origen de los errores de nuestro corpus. La mayoría de los errores considerados como etiológicamente ambiguos, se relacionan con los artículos (41 concordancias, un 2,9 % del total de errores, apareciendo cada 827,2 palabras) y con los verbos finitos simples (22 concordancias, conformando un 1,5 % del total y con un valor normalizado de 1.541,6).

De los procesos intralingüísticos posibles, la simplificación y la generalización siguen estando a la cabeza, destacando por encima del resto. Están muy presentes en todos los niveles, incluso en los iniciales. En el caso de la simplificación, se encuentran, sobre todo, en relación con el uso y formación de los verbos (81 concordancias, que conforman un 5,8 % del conjunto total de errores, y un valor normalizado de 418,7), los adjetivos (40 errores, que conforman un 2,85 % del total, y con valor normalizado de 847,8) y los artículos indefinidos (43 errores, un 3,07 % del total y un valor normalizado de 788,7). En los dos últimos está presente en problemas de concordancia de género y número.

La hipergeneralización de reglas es más frecuente en relación al uso de las preposiciones (44 errores, un 3,14 % del total, y un valor normalizado de 770,7) y en menor medida, en formas verbales compuestas (19 concordancias, que conforman un 1,3 % del total, y un valor normalizado de 1785). En el caso de las preposiciones, como ya hemos comentado, es habitual que se utilice una misma preposición para todos los contextos, ignorando, de alguna manera, los usos especiales o distintos de la misma. En las formas verbales compuestas, la generalización de reglas es común en nuestro corpus en la elección inapropiada de auxiliar para las formas compuestas, en su gran mayoría, para el *passé composé*.

Otros procesos de orden intralingüístico como la analogía, la asociación cruzada, la ignorancia de restricciones de las reglas, la neutralización o la aplicación incompleta de estas tienen menor presencia, e incluso no aparecen en algunos de los niveles, como los de A1 y A2, ya que son procesos que requieren un conocimiento más profundo de la lengua meta. Aparecen en errores muy específicos, y no de forma recurrente con una determinada categoría gramatical. De hecho, el número de apariciones es bastante reducido como para considerarlo significativo para el conjunto de nuestros aprendientes.

No obstante, detallaremos que la asociación cruzada en nuestro corpus aparece en primer lugar con formas verbales compuestas (6 concordancias, un 0,42 % del total de errores, y un valor normalizado de 5652,5). La analogía es más habitual con sustantivos (5 concordancias, conformando un 0,35 % del conjunto de errores, con un valor normalizado de 6783), provocando formas erróneas o incorrectas al establecer una comparación falsa. La aplicación de reglas incompletas aparece en formas verbales compuestas (5 errores, conformando un 0,35

% del conjunto de errores, con un valor normalizado de 6783) y la ignorancia de las restricciones de estas se reparte entre formas verbales compuestas, pronombres personales y artículos contractos (todas con igual número de concordancias: 5).

Finalmente, podemos destacar que se cumplen las expectativas respecto del origen de los errores de nuestro corpus, ya que la literatura en adquisición de lenguas señala a los errores intralingüísticos como los más frecuentes en el proceso de aprendizaje. No obstante, en nuestro corpus, esta diferencia no es muy grande, por lo que no podríamos confirmar las predicciones habituales. Los errores de origen ambiguo o desconocido, cuyo número es elevado en nuestro corpus, nos darían quizá un mayor margen entre ambos. Sin embargo, no podemos dejar de obviar que este análisis es el más subjetivo de todos, por lo que sería conveniente una revisión por distintos investigadores para que la definición del origen de los errores se asignara de la forma más objetiva posible.

3.4 Clasificación general de errores frecuentes

Como sabemos, nuestro AE se basa en la observación del mismo error desde los tres criterios antes descritos, de forma que podamos explicar completamente un error.

Así si unimos la categorización de los errores desde los criterios lingüístico, descriptivo y etiológico, la lista de los veinte tipos de errores más frecuentes en todos los niveles del corpus CORAF sería la siguiente[121]:

[121] Elegimos las veinte primeras categorías ya que tienen un número de apariciones significativo. Pese a categorizar 1400 errores, no tienen una tipología recurrente, sino que se dividen en gran variedad de tipos de error, sobre todo, por la categorización etiológica. Existen, por tanto, errores muy específicos, relacionados con cada aprendiente. Aún así, las 30 primeras categorías de error más frecuentes, con una frecuencia de apariciones igual y superior a diez, recogen el 57% de ellos. La lista de las primeras 150 categorías puede consultarse en el apéndice B.

Tipo de error	Descripción	Nº Errores	% total	Valor Normalizado
<LING_LEVEL><X><MAN><TARGET_MOD><ERROR_DESC><OM><ETIOLOGY><INTER><IFL1>	Error sintáctico por la omisión de un elemento sintáctico necesario en la oración, debido a un error interlingüístico, por interferencia de la lengua materna. Los ejemplos más habituales se relación con el olvido del sujeto (generalmente, un pronombre personal) y la segunda partícula de la negación (*pas*). Ej: **Malheureusement je n'ai beaucoup de temps* (C2M01) **Je ne suis très bonne avec le poisson* (A2W01) **est tragique* (B1W01)	99	7'07 %	342,5
<LING_LEVEL><L><SIG><TARGET_MOD><ERROR_DESC><MIF><ETIOLOGY><INTER><IFL1>	Error interlingüístico de léxico, relacionado con una forma incorrecta o errónea por interferencia de la L1. Habitual en aquellas palabras que el aprendiente crea sirviéndose de los conocimientos de la L1. Ej: *Je dois le *contrater* (ADSL) (C1W01) *Je pense faire un *entreviste* (A1M01)	99	7'07 %	342,5
<LING_LEVEL><L><SIG><TARGET_MOD><ERROR_DESC><MIS><ETIOLOGY><INTER><IFL1>	Error interlingüístico de léxico al usar una forma con un significado en un contexto donde no es válido, debido a la interferencia de la L1. Suele ocurrir al usar palabras del mismo campo léxico o sinónimos donde no es posible y en la introducción de "falsos amigos". Ej: *Le vocabulaire est très pareil*	49	3,5 %	692,1

Tipo de error	Descripción	Nº Errores	% total	Valor Normalizado
	(C1M01) *Même si le français procède de la même langue* (B1W02) *L'année passée, je suis allé* (C1M02)			
<LING_LEVEL><AMB> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA> <SIM>	Error ambiguo que implica una formación errónea, generalmente por una tendencia a la simplificación. Es habitual en las formas en las que no tenemos la certeza de que se trata de un error concreto, en la mayoría de los casos por una deficiente/errónea pronunciación que puede esconder algún otro fenómeno más. Habitual en la concordancia de género del adjetivo femenino y en la distinción *un/une* y *le/les*. Ej: *La littérature français(e) ou anglais (e)* (B2W02) *Il y a une-un place* (A2W02) *Tout(es) les choses* (B1W04)	49	3,5 %	692,1
<LING_LEVEL><G><CLA> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTRA> <HIG>	Error gramatical intralingüístico producido al confundir (seleccionar de forma inapropiada) algún elemento perteneciente a una clase funcional por querer generalizar reglas ya adquiridas. Es muy frecuente en las preposiciones, donde una de ellas es utilizada para todos los contextos de uso similares. Ej: *Je travaille à la banque (A1W03)*	43	3,07 %	788,7

Tipo de error	Descripción	Nº Errores	% total	Valor Normalizado
	Seulement dans la bibliothèque (A2M01) *Et même les exercises sont dans Internet* (B1M01)			
<LING_LEVEL><G><GEN> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA> <SIM>	Error gramatical intralingüístico por simplificación al concordar el género. Habitual en pronunciaciones incompletas de adjetivos, por ejemplo. Contrariamente al error que aparece en segundo lugar de esta lista, en este caso apreciamos claramente de que se trata de un error de competencia. *L'année *prochain* (B2W01) *J'étudie philologie français* (B2W02) *J'ai visité le Tour Eiffel* (A2W02)	41	2,9 %	821,2
<LING_LEVEL><G><CLA> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTER><IFL1>	Error gramatical al seleccionar de una clase un elemento para utilizarlo en un contexto donde no es válido por interferencia de la L1. Suele aparecer también en la selección de preposiciones, por tendencia a traducir de forma literal el enunciado. Ej: *Pour exemple* (A1M01) *Je connais le français pour les études en lycée* (B1W01) *En la radio* (A1W01) *D'aller à la France* (C1M01)	41	2,9 %	821,2
<LING_LEVEL><G><GEN> <TARGET_MOD><ERR	Error gramatical al concordar en género por	38	2,7 %	892,5

Tipo de error	Descripción	Nº Errores	% total	Valor Normalizado
OR_DESC><MIF> <ETIOLOGY><UNK>	causa que se desconoce, o que es ambigua para el investigador, pudiendo englobar varios procedimientos intralingüísticos. Ej: *Surtout faire des amis *françaises* (B2M01) *C'est un bon idée observer* (C2M01) *Je crois que c'est française (Daft Punk, groupe)* (A2W01)			
<LING_LEVEL><X><M AN> <TARGET_MOD><ERR OR_DESC><OM> <ETIOLOGY><INTRA> <SIM>	Error sintáctico por la omisión de un elemento de la oración debido a una simplificación del sistema lingüístico empleado. Suele aparecer en olvidos de partículas de la negación en aprendientes que, sin embargo, han podido utilizar bien la negativa antes, con lo cual no tenemos constancia de una completa adquisición por su parte, o en aprendientes de niveles inferiores que no elaboran su enunciado. Ej: **Je ne fais une escalade très difficile* (B1W01) *Comme mon mari a besoin de violon* (B1W02) *On a coupé* –L'internet au travail- (C1W01)	37	2,6 %	916,6
<LING_LEVEL><G><GE N> <TARGET_MOD><ERR OR_DESC><MIF> <ETIOLOGY><INTER>< IFL1>	Error gramatical de concordancia de género dando como resultado una forma errónea al interferir de forma negativa la L1. Habitual en formas donde	33	2,3 %	1.027,7

Tipo de error	Descripción	Nº Errores	% total	Valor Normalizado
	el género gramatical se altera al no coincidir con el género habitual de estos en la L1. Ej: **À la web* (B2M02) **Une certaine âge* (C2W02) *J'étais dans un équipe de basket* (B2W03)			
<LING_LEVEL><X><ORD> <TARGET_MOD><ERROR_DESC><WRO> <ETIOLOGY><INTER><IFL1>	Error sintáctico por el orden de los constituyentes de la oración por una interferencia de la L1. Es habitual en el orden de los adverbios con determinados tiempos verbales como el *passé composé.* Ej: *Il fait très bien tout* (A1W03) *Et je ne peux faire rien* (A2M01) *Et j'aime voyager beaucoup* (B1M01)	32	2,2 %	1.059,8
<LING_LEVEL><PHO> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA> <IGN>	Error fonético al producirse de forma errónea un enunciado por una causa intralingüística, al ignorar ciertas restricciones de las reglas. Aludimos aquí principalmente a problemas con las *élisions* obligatorias, que si bien son propias de la escritura, deberían de reflejarse también en la pronunciación, como las del pronombre '*je'* y el verbo conjugado o del pronombre '*que*'. Ej: *Mais j'avais étudié que un peu* (B2W02) **Parce que je aime comme les mots sonnent* (A2W02)	20	1,4 %	1.695,7

Tipo de error	Descripción	Nº Errores	% total	Valor Normalizado
<LING_LEVEL><G><PER><TARGET_MOD><ERROR_DESC><MIF><ETIOLOGY><INTER><IFL1>	Error gramatical al confundir la persona por influencia negativa de la L1. Habitual en aprendientes de niveles inferiores en los que usan una fórmula intermedia entre *il/elle* para todos los enunciados, muy similar a la española. Ej: **Elle est beau (son mari)* (A2W02) *Ils sont un peu fermés à son *circle d'amis* (C2W03)	17	1,21 %	1.995
<LING_LEVEL><G><TPS><TARGET_MOD><ERROR_DESC><MIF><ETIOLOGY><INTRA><SIM>	Error gramatical por formar erróneamente el tiempo verbal debido a un error intralingüístico, por simplificación del sistema de la interlengua. Ej: **Je faire exercice (A1M01) *Je vas dire toujours la même chose (quand elle travaillait à l'aéroport)* (C2W02)	16	1,1 %	2.119,7
<LING_LEVEL><G><CLA><TARGET_MOD><ERROR_DESC><MIS><ETIOLOGY><UNK>	Error gramatical de causa desconocida o, más frecuentemente, ambigua en la selección de un elemento de una clase funcional, no válido para un determinado contexto. Este error es frecuente en la selección y uso de preposiciones o conjunciones en contextos en los que no es adecuada, perteneciendo incluso a una categoría gramatical distinta. Ej: *Un petit bouquin qui s'appelle, *qui j'ai acheté à Paris* (C1W01)	16	1,1 %	2.119,7

Tipo de error	Descripción	Nº Errores	% total	Valor Normalizado
	*Je sais le nom *à espagnol* (B1W01)			
<LING_LEVEL><G><TPS><TARGET_MOD><ERROR_DESC><MIS><ETIOLOGY><INTRA><SIM>	Error gramatical en la selección del tiempo verbal por la simplificación del discurso. Suele aparecer en matices aspectuales de las formas verbales. Los aprendientes utilizan el pasado, pero no el tiempo del pasado que sería necesario, por ejemplo. La más corriente es la distinción entre el *imparfait* y el *passé composé*, y la del pasado y presente, por concordancia de tiempos. Ej: *J'avais l'opportunité de mettre en contact* (C1W03) *J'habite à Ceuta et je voyage beaucoup* (avant) (A1W02) *J'étudie* tourisme (A1W03)	15	1,07 %	2261
<LING_LEVEL><G><NBR><TARGET_MOD><ERROR_DESC><MIF><ETIOLOGY><INTER><IFL1>	Error gramatical al concordar en número, produciendo una forma errónea por interferencia de la L1. Sucede generalmente por la traducción literal de palabras en la L1, que tienen además, un número distinto en ambas lenguas. Ej: *La pâte* (A1M02) en vez de '*Les pâtes'.* **La gens est en France…* (C1M01) **Je trouve que le gens est poli* (C2M02)	15	1,07 %	2261
<LING_LEVEL><M><MFC>	Error de morfología flexiva	14	1 %	2422,5

Tipo de error	Descripción	Nº Errores	% total	Valor Normalizado
<TARGET_MOD><ERROR_DESC><MIF><ETIOLOGY><INTRA><SIM>	que tiene como consecuencia una forma errónea, producida por una simplificación de los elementos lingüísticos utilizados en el discurso. Es conveniente señalar que en CORAF está ligado a la conjugación del verbo '*étudier*.' Ej: **J'étude ici* (C2M02) *Ils *étudent* (A1M02) *Si on l'*étude* (C1M01)			
<LING_LEVEL><G><CLA><TARGET_MOD><ERROR_DESC><MIF><ETIOLOGY><INTER><IFL1>	Error gramatical por el uso de una forma perteneciente a una clase sintáctica inadecuada por interferencia de la L1. Ej: *Pour que j'adore comme travaille Dépardieu* (C1W01) *Je ne sais pas comme qualifier* (C2M02) *Que s'appelle Mxxx* (B1W03)	14	1 %	2422,5

Tabla 18: Resumen de las veinte categorías de errores más frecuentes en el corpus CORAF.

Como podemos observar, existe un primer problema en la categorización exhaustiva de errores. Al contar con numerosas categorías y subcategorías, encontramos errores muy específicos, lo que dispersa los errores en variadas tipologías. Quizá si abarcásemos las categorizaciones generales, y no buscásemos un nivel de detalle tan grande podríamos encontrar frecuencias de aparición mucho mayores. Así, si desde el punto de vista etiológico, nuestra distinción hubiera sido sólo entre interlingüístico e intralingüístico y ambiguo, la frecuencia de errores en cada uno de los tipos hubiera sido superior.

Por otra parte, la mayoría de los errores tienen un carácter gramatical, destacando problemas en la selección de elementos pertenecientes a clases cerradas o funcionales y en la concordancia del

género y número. En general, es habitual que existan problemas en la selección de categorías cerradas, ya que son elementos sintácticos que poseen muchas funciones distintas, y que además, no poseen una forma transparente, que ayude a la asignación de un determinado significado, lo que repercute en una gran ambigüedad a la hora de utilizarlos. Especialmente significativo en este tipo de error es el uso de las preposiciones, que tienden a ser neutralizadas en contextos similares, generalizando su uso.

No hemos encontrado, sin embargo, demasiados errores relativos al empleo de conjunciones, ya que, desgraciadamente, la frecuencia de uso de estas es mínima en nuestro corpus.

El hecho de que otro de los errores más frecuentes se relacione con la concordancia del género, ha de observarse, sobre todo, siendo conscientes de que el género femenino en el francés requiere una pronunciación particular, que, en muchas ocasiones, por simplificación del discurso, tiende a omitirse. Aunque en muchos de los casos se trata de un error claro de competencia, como hemos señalado, encontramos otros donde el error es ambiguo y no nos permite dilucidar si el aprendiente es consciente de que está utilizando la forma femenina o no. Evidentemente, es posible que en una producción escrita, el mismo aprendiente, por el tiempo de reflexión del que dispone y por la visualización de la palabra, no olvide la concordancia.

También encontramos errores sintácticos relativos a la omisión de elementos necesarios de la oración o por el contrario, a una alteración de su orden canónico. Entre los casos más comunes encontramos la omisión del sujeto (habitualmente representada por un pronombre personal) y una de las partículas necesarias para la construcción de un enunciado negativo (*pas*).

Así, los errores relativos al léxico se relacionan en su conjunto con el significado de los mismos, aunque, como ya indicamos en la explicación de nuestra taxonomía, se incluyen también aquí los relativos a un defecto de forma, ya que Granger (2003a) no lo señala entre sus categorías, ya que lo agrupa bajo los errores de ortografía (por tratarse de un análisis de errores en corpus de producción escrita)[122].

En nuestros resultados apreciamos tanto errores léxicos relacionados con la semántica, que suponen una inadecuada selección del término en el contexto de uso, muy en relación con el uso de *falsos amigos*

[122] La única explicación que lo diferencia es el criterio descriptivo, puesto que para las formas erróneas será marcado con la etiqueta <MIF> (forma errónea) y para los matices de carácter semántico, y siempre que no incurra también en una forma errónea, con la etiqueta <MIS>, de falsa selección.

y palabras del mismo campo léxico que no son sinónimas o intercambiables, y por otra parte, las relativas a la forma errónea de la palabra, muy en relación con fenómenos como el de la creación léxica o la traducción literal de palabras de la L1.

Pese a todos los tipos de errores que se han señalado, seguimos insistiendo en la idea de que el uso de la taxonomía de Granger (2003a) puede resultar algo ambigua, ya que, en muchas ocasiones los errores se interrelacionan, y un error sintáctico lo es también gramatical, o un error de léxico puede tener una estrecha relación con aspectos morfológicos. Y lo mismo ocurre con los subtipos dentro de cada categoría (gramática, sintaxis, léxico, morfología), ya que es muy difícil reflejar todos los errores que pueden producirse y el hecho de asignar sólo aquellos que están etiquetados, restringe mucho la explicación, además de suponer un grave problema en muchos de ellos, que bien pueden ser categorizados bajo varias de ellas, o bien no casan en ninguna de las expuestas.

3.5 Distribución de los errores generales frecuentes por niveles representados en CORAF

Los tipos de error antes descritos no se reparten de la misma forma en todos los niveles del MCER expresados en CORAF. Creemos así interesante establecer una comparativa de los diez primeros tipos de error y estudiar si se siguen manteniendo en todos los niveles, o si por el contrario, como resultaría de esperar, serían menos significativos (pese a que el número de palabras mayor de los niveles más avanzados supone un mayor número de errores).

Por consiguiente, mostraremos cada uno de los diez primeros tipos de error, estableciendo una gráfica que permita observar su frecuencia por niveles:

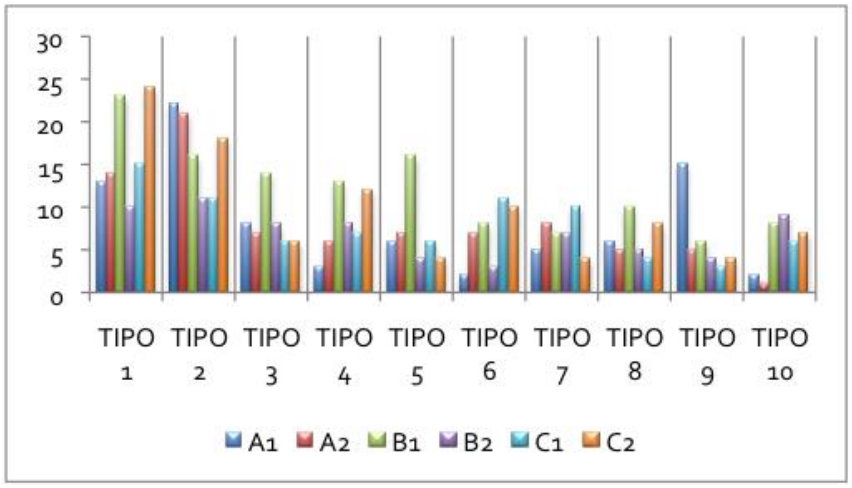

Gráfico 22: Comparativa de los diez errores del corpus CORAF más frecuentes por niveles del MCER.

ERROR	DESCRIPCIÓN
TIPO 1	Error sintáctico al omitir un elemento por interferencia de la L1.
TIPO 2	Error léxico por el uso de una forma incorrecta o errónea por interferencia de la L1.
TIPO 3	Error léxico al usar una forma en un contexto o con un significado donde no es válido, por interferencia de la L1.
TIPO 4	Error ambiguo debido a una formación errónea, generalmente por una tendencia a la simplificación. Formas donde puede categorizarse el error en más de un tipo.
TIPO 5	Error gramatical al seleccionar de forma inapropiada algún elemento perteneciente a una clase funcional por hipergeneralización.
TIPO 6	Error gramatical por simplificación al concordar el género.
TIPO 7	Error gramatical al seleccionar de una clase un elemento para utilizarlo en un contexto donde no es válido por interferencia de la L1.
TIPO 8	Error gramatical al concordar en género por causa que se desconoce, o que es ambigua para el investigador.
TIPO 9	Error sintáctico al omitir un elemento por la simplificación del sistema lingüístico empleado.
TIPO 10	Error gramatical en la concordancia de género dando como resultado una forma errónea por interferencia de la L1.

Tabla 19: Tabla explicativa de los diez tipos de errores más frecuentes.

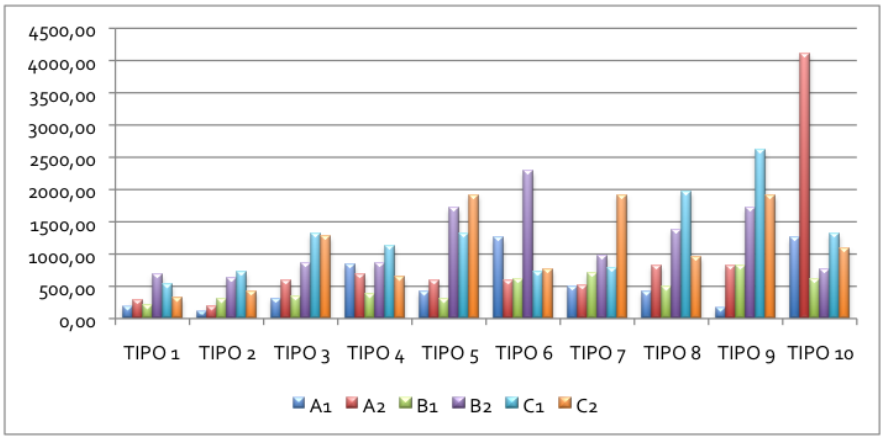

Gráfico 23: Diez tipos de error más frecuentes en valores normalizados para todos los niveles.

Si analizamos este gráfico, podemos observar que los niveles intermedio y avanzado (B2, C1, C2) son los más regulares en su comportamiento, obteniendo las cifras menos elevadas de errores en prácticamente todos los tipos. Sin embargo, sí que observamos algunos de ellos donde el nivel C2 ostenta un comportamiento poco habitual, ya que la experiencia nos dice que este tipo de aprendientes debería de tener menos frecuencias de errores en todas las categorías estudiadas. En este caso, no podemos más que acudir a las características especiales de los aprendientes de nivel C2 del corpus CORAF, ya que sus especificidades tienen que ver mucho en el resultado obtenido. Sin duda, encontramos aprendientes que no han adquirido el nivel (al realizar las grabaciones antes de la finalización del curso y quizá en niveles donde no se opta realmente a una adquisición de C2) y que además, tienen ciertos errores fosilizados que aparecen de forma recurrente, lo que altera la frecuencia de aparición.

De hecho, algunos de los tipos se ven incrementados por el discurso de un solo aprendiente, que comete el error de forma muy repetida. Por lo tanto, es necesaria una reformulación de nuestro AE, intentando, primero, contar con aprendientes que hayan verdaderamente adquirido el nivel C2, y segundo, un número mayor de sujetos, para poder generalizar y extrapolar nuestros resultados.

Finalmente, cabe destacar que no podemos, de alguna manera, realizar un análisis comparativo del todo fehaciente del desarrollo en la adquisición de la interlengua, ya que nuestro corpus no es de tipo longitudinal (aunque, como ya citamos, Granger, señala que este tipo de corpus es asimilable a este, ya que por sus características, y al recoger muestras de hablantes de todos los niveles en un determinado momento de la adquisición, es considerado *quasi/pseudo-longitudinal*).

3.5.1 Nivel A1: Usuario básico (Acceso)

Para el nivel A1, los errores más frecuentes aparecen en los tipos 2 y 9, estando fuertemente influenciados por la comparación con la L1. En el resto de tipos se mantiene en un discreto segundo plano, debido, sin duda, a la escasez y a la sencillez de las palabras producidas por los aprendientes analizados. Son formas muy básicas, muchas de ellas palabras sueltas y monosílabos, que suelen repetir las construcciones previamente dadas por el contexto de estudio/interlocutor, por lo que el uso real de la lengua meta no es aún tan desarrollado, lo que sin duda, repercute en un menor número de errores.

Entre las dificultades de los aprendientes de nivel A1 del corpus CORAF estarían, por tanto, el léxico, produciendo formas erróneas por influencia de la L1, como en *'*entreviste*' (A1M01), '*la *gramatique'* (A1W03) o '*l'*ambiante'* (A1W01), y también de sintaxis, omitiendo elementos necesarios de la oración por procedimientos de simplificación. Algunos ejemplos podrían ser: '*à Nancy les gens plus sympatiques'* (A1W02); '*c'est près ici'* (A1W03); '*jouer basket'* (A1M02); '*a beaucoup de monuments*' (A1W01).

Si lo comparamos con los datos obtenidos para su nivel, las diez primeras categorías con más errores son las siguientes:

Errores	Nº	% total A1	Valor Normalizado	TIPO
<LING_LEVEL><L><SIG> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	22	10,8%	113,9	TIPO 2

Errores	Nº	% total A1	Valor Normalizado	TIPO
<LING_LEVEL><X><MAN> <TARGET_MOD><ERROR_DESC><OM> <ETIOLOGY><INTRA><SIM>	15	7,3 %	167	TIPO 9
<LING_LEVEL><X><MAN> <TARGET_MOD><ERROR_DESC><OM> <ETIOLOGY><INTER><IFL1>	13	6,4 %	192,7	TIPO 1
<LING_LEVEL><G><TPS> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTRA><SIM>	11	5,4 %	227,8	
<LING_LEVEL><L><SIG> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTER><IFL1>	8	3,9 %	313,2	TIPO 3
<LING_LEVEL><G><CLA> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTRA><HIG>	6	2,9 %	417,6	TIPO 5
<LING_LEVEL><G><GEN> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><UNK>	6	2,9 %	417,6	TIPO 8
<LING_LEVEL><G><CLA> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTRA><SIM>	5	2,4 %	501,2	
<LING_LEVEL><G><CLA> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTER><IFL1>	5	2,4 %	501,2	TIPO 7
<LING_LEVEL><L><SIG> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTRA><ASC>	4	1,9 %	626,5	

Tabla 20: Tipos de error frecuentes para el nivel A1 del corpus CORAF.

El nivel A1 sigue la tendencia general, aunque no en el mismo orden, e introduce un error que se sitúa dentro de los veinte más frecuentes para todos los niveles, que alude a la distinción de tiempos verbales, y que se relaciona, sobre todo, con el uso incorrecto del presente y pasado, que aún no ha sido adquirido, como vemos en los siguientes ejemplos:

'mes parents travaillent ici pendant le protectorat' (A1W02)

'quand j'ai deux ans, il sont venus…' (A1W03)

'mais je ne le fini pas' (A1W03)

'je faire exercice' (A1M01)

3.5.2 Nivel A2: Usuario básico (plataforma)

El nivel A2 tiene mayor dificultad en el tipo 2 (formación errónea del léxico por interferencia de la lengua materna), manteniéndose en el resto. El léxico sigue siendo uno de los principales problemas, como ocurre con el nivel A1. Sin embargo, destaca en el tipo 10, relativo a problemas de concordancia de género por la interferencia de la L1, dando como resultado una forma errónea, donde apenas encontramos un solo error.

De forma más detallada, podemos observar la lista de frecuencias de error más usuales para dicho nivel:

Errores	Nº	% total A2	Valor Normalizado	TIPO
<LING_LEVEL><L><SIG> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	21	10,1 %	195,7	TIPO 2
<LING_LEVEL><X><MAN> <TARGET_MOD><ERROR_DESC><OM> <ETIOLOGY><INTER><IFL1>	14	6,7 %	293,5	TIPO 1
<LING_LEVEL><G><CLA> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTER><IFL1>	8	3,8 %	513,7	TIPO 7
<LING_LEVEL><L><SIG> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTER><IFL1>	7	3,3 %	587,1	TIPO 3
<LING_LEVEL><G><CLA> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTRA><HIG>	7	3,3 %	587,1	TIPO 5
<LING_LEVEL><G><GEN> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	7	3,3 %	587,1	TIPO 6
<LING_LEVEL><G><AUX> <TARGET_MOD><ERROR_DESC><MIF>	6	2,9 %	685	

Errores	Nº	% total A2	Valor Normalizado	TIPO
<ETIOLOGY><INTRA><HIG>				
<LING_LEVEL><AMB> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	6	2,9 %	685	TIPO 4
<LING_LEVEL><X><MAN> <TARGET_MOD><ERROR_DESC><OM> <ETIOLOGY><INTRA><SIM>	5	2,4 %	822	TIPO 9
<LING_LEVEL><X><ORD> <TARGET_MOD><ERROR_DESC><WRO> <ETIOLOGY><INTER><IFL1>	5	2,4 %	822	

Tabla 21: Tipos de error frecuentes para el nivel A2 del corpus CORAF.

Como ya hemos comentado, el nivel A2 sigue la tónica general, aunque introduce dentro de los más frecuentes dos nuevos errores al ir ampliando más su discurso:

- Error sintáctico al cambiar el orden canónico de la oración por influencia de la L1

  - **Je ne peux comprendre pas* (A2M01)
  - **Comment est-ce que se dit ça?* (A2M02)
  - *je ne pas savais* (A2W02)

- Error gramatical en el auxiliar del verbo principal al conjugar tiempos compuestos[123]:

  - *Et quand j'ai allé* (A2W01)
  - *C'est pour ça que je m'ai marié avec lui* (A2W02)

[123] Esta nueva categoría es más frecuente porque el aprendiente A2W01 repite insistentemente este error en su narración.

3.5.3 Nivel B1: Usuario independiente (umbral)

El nivel B1 en nuestro corpus cuenta con una cierta irregularidad en la frecuencia de sus errores, situándose no en pocas ocasiones a la cabeza de los errores cometidos para los tipos estudiados. Así, el nivel B1 tiene generalmente dificultades en los tipos 1, 3, 4, 5 y 8, y junto con el B2, también para el tipo 10. El discurso se va progresivamente complicando más y son, por tanto, más patentes los errores conocidos.

Destaca por encima de todos el tipo 1, de orden sintáctico, por la omisión de constituyentes de la oración debido a la interferencia de la L1, y que, como veremos, aumenta también en el nivel C2 de forma muy significativa.

Se producen problemas también en el léxico, pero ya no tanto de orden formal, sino de orden semántico, utilizando vocabulario con un significado en contextos donde no se permite (tipo 3). Es el caso de la utilización cada vez más recurrente de *falsos amigos*, de confusión entre palabras del mismo campo semántico (utilización de léxico inadecuado) o de problemas en el régimen de complementos verbales y otras colocaciones. Lo podemos observar en ejemplos como los siguientes:

Il étudie médecin (B1W01)

Tu peux retourner dans le même jour (B1W02)

C'est un site que j'aime beaucoup (B1W04)

Aumenta además el número de errores ambiguos (tipo 4), por una pronunciación deficiente o incompleta, junto con la utilización de los elementos de forma correcta no en todas las ocasiones, lo que supone un problema para el investigador, que bien, encuentra errores que pueden categorizarse en varios tipos a la vez, o que tiene serias dudas para dilucidar si se trata sólo de un error fonético o si este conlleva también fallos en la competencia. Algunos ejemplos que podemos observar son los siguientes:

Dans <u>un</u> résidence (B1W03): ¿Error de pronunciación, de forma del determinante indefinido o de concordancia de género?

<u>Tout(s)</u> les choses (B1W04): ¿Error de pronunciación o de concordancia en género y número?

Je ne sais pas beaucoup de choses de la musique français (B1M01): ¿error de pronunciación o de concordancia de género?

Y por otro lado, es más patente el número de errores en la utilización de elementos de clases funcionales por una hipergeneralización (tipo 5), siendo muy habitual en la selección de las preposiciones:

Il est dans l'école (B1W01)

La fac d'histoire en Cxxxx (B1W02)

J'habite aussi en Axxx (B1W03)

Vemos así que el aprendiente está poniendo también en juego nuevos elementos y procesos para la comunicación, como es el caso de los procesos internos de comparación entre formas, reglas y estructuras propias de la lengua meta, apareciendo ya errores que tienen su origen en la simplificación y la generalización de reglas conocidas.

Si observamos la lista de categorías más frecuentes para el nivel B1, nos damos cuenta de que refleja todos los tipos de error estudiados, aunque variando en el orden de su frecuencia:

Errores	Nº	% total B1	Valor Normalizado	TIPO
<LING_LEVEL><X><MAN> <TARGET_MOD><ERROR_DESC><OM> <ETIOLOGY><INTER><IFL1>	23	8,2 %	213,3	TIPO 1
<LING_LEVEL><L><SIG> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	16	5,7 %	306,7	TIPO 2
<LING_LEVEL><G><CLA> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTRA><HIG>	16	5,7 %	306,7	TIPO 5
<LING_LEVEL><L><SIG> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTER><IFL1>	14	5 %	350,5	TIPO 3
<LING_LEVEL><AMB> <TARGET_MOD><ERROR_DESC><MIF>	13	4,6 %	377,5	TIPO 4

Errores	Nº	% total B1	Valor Normalizado	TIPO
<ETIOLOGY><INTRA><SIM>				
<LING_LEVEL><G><GEN> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><UNK>	10	3,5 %	490,8	TIPO 8
<LING_LEVEL><G><GEN> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	8	2,8 %	613,5	TIPO 10
<LING_LEVEL><G><GEN> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	8	2,8 %	613,5	TIPO 6
<LING_LEVEL><G><CLA> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTER><IFL1>	7	2,5 %	701,1	TIPO 7
<LING_LEVEL><X><MAN> <TARGET_MOD><ERROR_DESC><OM> <ETIOLOGY><INTRA><SIM>	6	2,1 %	818	TIPO 9

Tabla 22: Lista de categorías frecuentes de error para el nivel B1 en CORAF.

3.5.4 Nivel B2: usuario independiente avanzado

El nivel B2 es el que mantiene un desarrollo más regular, alcanzado un número de errores bastante reducido en la práctica totalidad de los tipos estudiados. No obstante, no significa que esté exento de dificultades, pero sí que es el nivel que tiene menor número de errores en muchos de ellos.

De las categorías más frecuentes, el nivel B2 obtiene un peor resultado en el tipo 10, relacionado con errores gramaticales de concordancia de género por interferencia de la L1. A medida que se va utilizando una mayor cantidad de vocabulario, de campos léxicos variados, se van produciendo nuevos errores al no tener completamente adquiridas todas las formas y sus características. Podemos observar algunos ejemplos como:

C'est le limite plus bas (B2M01)

C'est une film espagnol (B2W01)

Je crois que c'est un bon méthode (B2M02)

Entre los errores más frecuentes para el nivel B2 se sitúan:

Errores	Nº	% total B2	Valor Normalizado	TIPO
<LING_LEVEL><L><SIG> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	11	5,8 %	623,45	TIPO 2
<LING_LEVEL><X><MAN> <TARGET_MOD><ERROR_DESC><OM> <ETIOLOGY><INTER><IFL1>	10	5,2 %	685,8	TIPO 1
<LING_LEVEL><PHO> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><IGN>	10	5,2 %	685,8	
<LING_LEVEL><G><GEN> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	9	4,7 %	762	TIPO 10
<LING_LEVEL><L><SIG> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTER><IFL1>	8	4,2 %	857,2	TIPO 3
<LING_LEVEL><AMB> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	8	4,2 %	857,2	TIPO 4
<LING_LEVEL><G><CLA> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTER><IFL1>	7	3,7 %	979,7	TIPO 7
<LING_LEVEL><G><GEN> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><UNK>	5	2,6 %	1.371,6	TIPO 8
<LING_LEVEL><PHO> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	4	2,1 %	1.714,5	
<LING_LEVEL><G><CLA> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTRA><HIG>	4	2,1 %	1.714,5	TIPO 5

Tabla 23: Lista de categorías de errores frecuentes para el nivel B2.

El nivel B2 sigue las tendencias habituales para el conjunto de niveles, pero introduce ya nuevas categorías, sobre todo, relacionadas con errores de orden fonético. Por un lado, aquellos debidos a una simplificación de la forma que conduce a una pronunciación incorrecta o incompleta, y por otro, errores que se producen por ignorar ciertas restricciones en las reglas. En este caso concreto, incluimos aquí las llamadas *élisions* obligatorias[124], que como ya hemos explicado anteriormente, no se pueden señalar de ninguna otra forma, ya que la taxonomía de Granger (2003a) está concebida para corpus escritos, y refleja este aspecto en la parte relativa a la ortografía. Algunos ejemplos de esta tendencia son los siguientes:

Mais j'avais étudié que un peu (B2W02)

Je xxx de la utiliser (B2M02)

3.5.5 Nivel C1: Usuario competente: dominio operativo eficaz

El nivel C1 mantiene también un desarrollo bastante regular, con errores en todos los tipos, pero prácticamente sin alcanzar grandes frecuencias en muchos de ellos, salvo para los tipos 6 y 7, errores gramaticales de concordancia del género por simplificación y de falsa selección en clases funcionales por interferencia de la L1, respectivamente.

Observamos así la lista de frecuencias de los tipos de error más usuales de los aprendientes de este nivel en CORAF:

Errores	Nº	% total C1	Valor Normalizado	TIPO
<LING_LEVEL><X><MAN><TARGET_MOD><ERROR_DESC><OM><ETIOLOGY><INTER><IFL1>	15	5,9 %	524,5	TIPO 1
<LING_LEVEL><L><SIG><TARGET_MOD><ERROR_DESC><MIF><ETIOLOGY><INTER><IFL1>	11	4,3 %	715,2	TIPO 2

[124] La aparición de este fenómeno está muy ligada a la producción de un determinado aprendiente, B2W02, que produce la mayoría de los errores.

Errores	Nº	% total C1	Valor Normalizado	TIPO
<LING_LEVEL><G><GEN> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	11	4,3 %	715,2	TIPO 6
<LING_LEVEL><G><CLA> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTER><IFL1>	10	3,9 %	786,8	TIPO 7
<LING_LEVEL><AMB> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	7	2,7 %	1.124	TIPO 4
<LING_LEVEL><G><GEN> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	6	2,3 %	1.311,3	TIPO 10
<LING_LEVEL><PHO> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><IGN>	6	2,3 %	1.311,3	
<LING_LEVEL><G><CLA> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTRA><HIG>	6	2,3 %	1.311,3	TIPO 5
<LING_LEVEL><L><SIG> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTER><IFL1>	6	2,3 %	1.311,3	TIPO 3
<LING_LEVEL><G><CLA> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><UNK>	5	1,9 %	1.573	

Tabla 24: Tipos de errores frecuentes del nivel C1 en el corpus CORAF.

En el nivel C1 podemos ver que se mantienen los tipos generales, aunque se introducen dos nuevos, uno ya mencionado anteriormente para el nivel B2, como es el error fonético por ignorancia de las restricciones de las reglas, y otro, que refleja la dificultad para seleccionar el elemento de clase funcional adecuado, pero cuya causa es desconocida o ambigua para el investigador.

Entre los primeros, asociados especialmente a dos de los aprendientes, lo que altera significativamente el número de apariciones de los mismos para el conjunto de hablantes de C1 en CORAF, podemos encontrar ejemplos como:

Et devant le ordinateur (C1W01)

De aller à la France (C1M01)

Entre los de orden gramatical y relativos a una falsa selección de elementos de clases cerradas, encontramos, entre otros ejemplos:

**Dès de troisième année* (C1M01)

*Un petit bouquin qui s'appelle, *qui j'ai acheté à Paris* (C1W01)

*Il y a un film *qui je pense c'est magnifique* (C1W01)

3.5.6 Nivel C2: usuario competente: maestría

El nivel C2, como ya hemos comentado anteriormente, tiene varios problemas asociados que impiden una generalización de sus resultados. Por un lado, hemos asumido que los aprendientes escolarizados en los niveles más avanzados (y finales) de los centros educativos visitados son los que podrían formar parte de este nivel. Por otro lado, el momento en el que se realizan las grabaciones, así como las especificidades de los aprendientes que se han prestado voluntarios, nos impiden pensar lo anterior.

No obstante, hemos seguido analizando este nivel junto con el resto, y aunque observamos que el porcentaje de errores es bastante similar a lo esperado, ya que para el total de palabras producidas no encontramos un número alarmante de errores, sí que se pone de manifiesto que nuestros aprendientes no han alcanzado el nivel esperado. De haberlo hecho, el porcentaje de errores en algunas categorías estudiadas debería de haber sido mucho menor (como las relativas a errores de léxico, o gramaticales de concordancia de género y número).

Así, y basándonos en los datos obtenidos, podemos ver que el nivel C2 suele ser el que menor número de errores contiene en muchos de los diez tipos estudiados. Sin embargo, encuentra graves dificultades en algunos de ellos, como el tipo 1 (error sintáctico por la omisión de un elemento necesario por la influencia de la L1), y junto con el nivel B2, en el tipo 6 (error gramatical de concordancia de género por la simplificación). Nos sorprende que exista aún una gran interferencia de la L1, como demuestra el tipo 1, aunque es quizá habitual si tenemos en

cuenta que muchos de nuestros aprendientes traducen de forma literal aquello que quieren decir, estando más centrados en el contenido que en la forma.

En cuanto al error sintáctico por omisión, creemos que es bastante habitual ya que al intentar mantener una conversación fluida, pueden olvidar algunos constituyentes por su preocupación en otros aspectos como la pronunciación, o adecuación al contenido (amén de la presión extra que supone una interacción oral con un interlocutor desconocido y que está siendo grabada).

Como para el resto de niveles estudiados, podemos observar también la lista de tipos de error más usual para este nivel:

Errores	Nº	% total C2	Valor Normalizado	TIPO
<LING_LEVEL><X><MAN> <TARGET_MOD><ERROR_DESC><OM> <ETIOLOGY><INTER><IFL1>	24	8, 5 %	319,4	TIPO 1
<LING_LEVEL><L><SIG> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	18	6,4 %	425,8	TIPO 2
<LING_LEVEL><AMB> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	12	4,2 %	638,8	TIPO 4
<LING_LEVEL><X><ORD> <TARGET_MOD><ERROR_DESC><WRO> <ETIOLOGY><INTER><IFL1>	11	3,9 %	696,9	
<LING_LEVEL><G><GEN> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	10	3,5 %	766,6	TIPO 6
<LING_LEVEL><G><GEN> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><UNK>	8	2,8 %	958,2	TIPO 8
<LING_LEVEL><G><PER> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	8	2,8 %	958,2	
<LING_LEVEL><G><GEN> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	7	2,4 %	1.095, 14	TIPO 10
<LING_LEVEL><G><NBR> <TARGET_MOD><ERROR_DESC><MIF>	7	2,4 %	1.095, 14	

Errores	Nº	% total C2	Valor Normalizado	TIPO
<ETIOLOGY><INTER><IFL1>				
<LING_LEVEL><L><SIG> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTER><IFL1>	6	2,1 %	1.277,6	TIPO 3

Tabla 25: Lista de tipos de error más frecuentes para el nivel C2 en CORAF.

El nivel C2 mantiene los tipos de error generales que hemos visto anteriormente, pero incluye un mayor número en relación con tres nuevas categorías, causadas por una interferencia de la L1.

El primero es un error sintáctico por una alteración del orden canónico de la oración, como se refleja en los siguientes ejemplos:

**Je n'ai retournée jamais* (C2W01)

*Presque on comprend tout ce qu'elle dit (*C2M02)

C'est ça que fait les jeunes/les gens ici (C2W03)

El segundo y el tercero están íntimamente relacionados y distinguen un error gramatical al confundir, por interferencia de la L1, la persona y/o el número de determinados constituyentes de la oración. Como veremos, la mayoría de ellos aluden a una sola forma: el sustantivo '*gens*', que causa un recurrente error al intentar realizar la traducción literal, puesto que el número es diferente en español y en francés. También hay que destacar que el número de errores aumenta porque está ligado al discurso de dos aprendientes en particular, que lo cometen en numerosas ocasiones. Algunos ejemplos de estas categorías son:

*Je trouve que *le gens est poli* (C2M02)

Le gens est un peu… (C2W03)

*Ils sont un peu fermés à son *cercle d'amis* (C2W03)

(À Paris) le gens ce n'est pas sympathiques (C2M02)

3.6 Tipología general de errores frecuentes en relación a la parte de la oración afectada

La lista de categorías de errores recurrentes es un elemento muy útil para distinguir dónde se encuentran las dificultades de los aprendientes. Sin embargo, esta lista puede ser mucho más útil si podemos identificar qué parte o constituyentes de la oración son los que contienen o suelen inducir a mayor número de errores. Por tanto, relacionar la lista de errores más frecuentes junto con la categoría afectada, puede mostrarnos aún más pistas sobre las necesidades específicas de nuestros aprendientes.

Para ello, añadimos a nuestro AE la taxonomía referente a la categoría gramatical, que se encuentra englobada dentro del criterio lingüístico. Así, si unimos el resto de criterios (lingüístico, descriptivo y etiológico) y las categorías afectadas, obtenemos la lista de los doce errores más usuales del corpus CORAF para todos los niveles del MCER[125]:

	Errores	Descripción	Nº	% total	Valor Normalizado
1	<LING_LEVEL><L><SIG><GRAM_CAT><NOUN><NOM><TARGET_MOD><ERROR_DESC><MIF><ETIOLOGY><INTER><IFL1>	Error léxico al formar un sustantivo de manera incorrecta por influencia de la L1. Habitual en fenómenos de creación léxica.	49	3,5 %	692,1
2	<LING_LEVEL><G><CLA><GRAM_CAT><PREPOSITION><PES><TARGET_MOD><ERROR_DESC><MIS><ETIOLOGY><INTRA><HIG>	Error gramatical al seleccionar de forma incorrecta una preposición por una generalización inadecuada de las reglas ya adquiridas. Muy habitual con las preposiciones '*dans'* y '*à*', que se generalizan para todos los contextos.	40	2,8 %	847,9 %

[125] Elegimos los doce primeros tipos de error para establecer una mejor comparación al tener un número de apariciones más alto. Nuestro corpus contiene errores muy específicos, teniendo la mayoría de ellos, una o dos apariciones en el conjunto de los aprendientes.

	Errores	Descripción	Nº	% total	Valor Normalizado
3	<LING_LEVEL><X><MAN><GRAM_CAT><PRONOUN><POO><TARGET_MOD><ERROR_DESC><OM><ETIOLOGY><INTER><IFL1>	Error sintáctico al omitir un pronombre personal en la oración por interferencia de la L1. Habitualmente se relaciona con la omisión del sujeto de la oración, o por la no inclusión de pronombres de COD/COI.	31	2,2 %	1.094
4	<LING_LEVEL><X><MAN><GRAM_CAT><ADVERB><ADV><TARGET_MOD><ERROR_DESC><OM><ETIOLOGY><INTER><IFL1>	Error sintáctico al omitir un adverbio por interferencia de la L1. En este caso concreto se trata de la formación incompleta de la negación, obviando la segunda parte, la partícula '*pas*'.	30	2,1 %	1.130,5
5	<LING_LEVEL><G><CLA><GRAM_CAT><PREPOSITION><PES><TARGET_MOD><ERROR_DESC><MIS><ETIOLOGY><INTER><IFL1>	Error gramatical al seleccionar de forma incorrecta una preposición por interferencia de la L1. Se produce generalmente por traducir de forma literal el enunciado.	26	1,8 %	1.304,4
6	<LING_LEVEL><L><SIG><GRAM_CAT><NOUN><NOM><TARGET_MOD><ERROR_DESC><MIS><ETIOLOGY><INTER><IFL1>	Error léxico al seleccionar de forma inadecuada un sustantivo, que aún siendo correcto, no puede utilizarse en dicho contexto. Todo ello causado por una interferencia de la L1. Habitual en la selección de palabras del mismo campo léxico de forma inadecuada (ej.: *place, endroit, lieu//cours, classe*) y en el uso de *falsos amigos*.	25	1,7 %	1.356,6
7	<LING_LEVEL><AMB><GRAM_CAT><ARTICLE><AIN><TARGET_MOD><ERROR_DESC><MIF><ETIOLOGY><INTRA><SIM>	Error ambiguo por utilización de la forma errónea del artículo indefinido por simplificación. En este caso, encontramos una pronunciación deficiente del mismo que no nos	21	1,5 %	1.615

	Errores	Descripción	Nº	% total	Valor Normalizado
		permite distinguir entre la forma para el femenino y la del masculino, no pudiendo distinguir, si además del error fonético existe otro gramatical por la concordancia del género.			
8	<LING_LEVEL><X><ORD><GRAM_CAT><ADVERB><ADV><TARGET_MOD><ERROR_DESC><WRO><ETIOLOGY><INTER><IFL1>	Error sintáctico por alterar la posición habitualmente establecida para un adverbio, causado por la interferencia de la L1.	18	1,2 %	1.884,1
9	<LING_LEVEL><AMB><GRAM_CAT><ADJECTIVE><ADJ><TARGET_MOD><ERROR_DESC><MIF><ETIOLOGY><INTRA><SIM>	Error ambiguo en el uso del adjetivo por una simplificación del sistema fonético. La pronunciación no nos permite distinguir si además del error fonético existe otro de tipo gramatical en la concordancia del género y/o número.	16	1,1 %	2.119,6
10	<LING_LEVEL><M><MFC><GRAM_CAT><VERB><VSC><TARGET_MOD><ERROR_DESC><MIF><ETIOLOGY><INTRA><SIM>	Error en la morfología flexiva de una forma verbal finita simple por simplificación.	14	1 %	2.422,5
11	<LING_LEVEL><G><GEN><GRAM_CAT><ARTICLE><ADE><TARGET_MOD><ERROR_DESC><MIF><ETIOLOGY><UNK>	Error gramatical de causa ambigua y desconocida al seleccionar de forma errónea el artículo definido. En ocasiones, se trata de la utilización del artículo '*le*' de forma indiscriminada.	14	1 %	2.422,5
12	<LING_LEVEL><G><AUX><GRAM_CAT><VERB><VCC>	Error gramatical por la utilización errónea del auxiliar en las formas verbales finitas	14	1 %	2.422,5

	Errores	Descripción	Nº	% total	Valor Normalizado
	<TARGET_MOD><ERRO R_DESC><MIF> <ETIOLOGY><INTRA>< HIG>	compuestas, causado por la generalización de las reglas ya adquiridas. El caso más común es el uso incorrecto del auxiliar en la formación del *passé composé*.			

Tabla 26: Errores más frecuentes y parte de la oración afectada en el corpus CORAF para todos los niveles del MCER.

Como podemos deducir de la anterior lista, los más numerosos son los de orden gramatical, relacionados con problemas en la selección y uso de las preposiciones y con menor presencia, del verbo auxiliar en formas verbales compuestas. En todos ellos, la causa más frecuente es de orden intralingüístico, por la simplificación y la (hiper)generalización de reglas conocidas. Podemos observar una muestra de ello en:

Je ne me souviens pas le nom dans ce moment (B2M01)

Je suis né dans Axxx (A1M01)

**Ma mère et mon père ont allé à Paris* (A1W03)

Entre los errores gramaticales, también encontramos problemas con el uso de artículo definido, por dificultades en la concordancia de género, cuya causa es ambigua o desconocida. Este error no es quizá representativo, y puede estar muy relacionado con la producción oral, ya que el artículo definido es uno de los primeros en adquirirse, siendo el indefinido el que más dificultades plantea para la adquisición (Cf. Véronique, 2009). Además, el número total de apariciones (sólo 14) es muy limitado para considerarlo representativo, y puede estar ligado al discurso de un solo aprendiente, que como en otras categorías, elevaría su número de apariciones. Podemos observar algunos ejemplos de este tipo en:

C'est le vrai forme (C1W01)

Et le région Nord-Pas-de-Calais (C1W01)

Le premier(e) fois que je vais à France (C2M01)

En segundo lugar, encontramos errores sintácticos por la interferencia de la L1, destacando la omisión del pronombre personal y del adverbio en la oración, y problemas en la posición de este en la oración.

El primero suele producirse por una omisión del sujeto, y en menor medida, del COD y del COI, sobre todo en posición preverbal. Algunos ejemplos característicos son:

**Je propose voir* (A1M02), (haciendo una proposición a la entrevistadora)

**Je crois que sont* (C1M01)

**sont sympathiques* (B2W02)

El adverbio, por su parte, suele ser omitido en la formación de la negación (a través de la partícula '*pas*') y alterado en su posición habitual en la oración, siendo muy frecuente con formas verbales compuestas. Podemos apreciar estas categorías en:

*Je pense que *je ne parle bien français* (C2M01)

**Je ne /ve/ les différences* (A1W01)

Toujours il y a quelque chose en français (C1W01)

J'ai aimé beaucoup Belgique (A2W02)

En tercer lugar encontramos errores relacionados con el léxico y el uso del sustantivo, tanto por su forma errónea como por su inadecuada selección para el contexto de uso, y en ambos casos, por la fuerte influencia de la L1. Los sustantivos con desviaciones en su forma son los más numerosos de todos los errores, y suelen estar relacionados con una deficiente competencia léxica. El aprendiente alude a referentes que no conoce o a léxico que aún no ha aprendido, pero necesita comunicar su mensaje y para ello, prueba y pone en práctica ciertas estrategias de comunicación que le llevan a inventar, manipular o cambiar formas que conoce, sobre todo de su L1. Es un proceso muy habitual que se produce, en la mayoría de las ocasiones, por una traducción literal de la L1, aunque no exenta de ciertos procedimientos intralingüísticos. Los

aprendientes no traducen la palabra sin más, sino que la dotan de apariencia próxima a la lengua meta. Sería un procedimiento similar al realizado para los calcos y los préstamos, pero teniendo en cuenta que las formas definitivas son incorrectas. Algunos ejemplos que hemos encontrado para demostrarlo son los siguientes:

*Et dans l'avenue de la *reconquiste* (A1M01)

*Et avec une *ambiante très bonne* (A2W01)

*La *gramatique* (A1W03)

*La *sutilesse* (C1M01)

Además, encontramos otro gran número de ellos referentes a una inadecuada selección del léxico, también causados por una fuerte influencia de la L1. En este caso, suelen tener una forma/apariencia correcta, pero su significado no lo es. Aludimos por tanto a problemas relacionados más con la semántica, ya que podemos ver palabras que pertenecen al mismo campo léxico, pero que no son intercambiables, en lo que denominamos un empleo de léxico inadecuado; y por otro lado, el uso de *falsos amigos*. Algunos ejemplos prototípicos pueden ser:

Tu peux rechercher l'Axxx et différents sites pour visiter (A1M01)

Je pense que j'entends bien (C2M01), en lugar de '*comprendre*'.

Si c'est pour la nuit… (A2W01), en lugar de '*soir*'.

*C'est *un population grand* (C1W01), en lugar de '*village*'.

Los errores morfológicos son los menos numerosos en la lista de los errores más frecuentes, afectando sólo a la morfología flexiva en verbos, por simplificación de las formas y en la aplicación de las reglas conocidas. En nuestro corpus va muy unido al verbo '*étudier*', que la práctica totalidad de los aprendientes no es capaz de producir de forma correcta en sus formas del presente de indicativo (quizá también influidos por la lengua materna, más tendente a la formación de palabras llanas). Algunos ejemplos posibles son:

*J'*étude ici* (C2M01)

*Je *rechesse des différents sites (sur internet) des langues étrangères* (A1M01)

*Si on l'*étude* (C1M01)

Mención aparte merecen dos tipos de errores cuya descripción lingüística es ambigua. Se relaciona con el uso del adjetivo y del artículo indefinido. La ambigüedad existe porque se produce una pronunciación incorrecta o incompleta, que no sabemos si enmascara, además, errores de tipo gramatical por dificultades en la concordancia del género. Se simplifica la pronunciación, lo que impide ver con claridad si el error es puramente fonético o por el contrario, mucho más complejo. Es evidente que estos errores se deben al tipo de interacción, puesto que la ambigüedad está marcada por la producción oral, y el mismo aprendiente, en un corpus escrito, podría no tener ningún problema al respecto.

Así, la pronunciación del artículo indefinido en su forma femenina es complejo por el fonema /y/, por lo que el aprendiente tiende a producirla, pero sin excesiva claridad, sobre todo si tenemos en cuenta que pretende mantener el flujo normal de la interacción. En otros, y para el caso del adjetivo, es habitual que el aprendiente otorgue al sustantivo el género correcto, pero olvide concordarlo. Podemos comprobarlo en los siguientes ejemplos:

Littérature français ou anglais (B2W02)

Il y a une personnage (A2W02)

J'aime bien la musique français (B1W03)

Ils ont un histoire (B2M02)

En general, la causa más usual de los errores, como podemos observar, está muy equilibrada, ya que tenemos aproximadamente el mismo número de errores interlingüísticos (por interferencia de la lengua materna) que intralingüísticos, algo que, por otra parte, es una constante en nuestro corpus, como ya observamos anteriormente en este capítulo.

3.7 Distribución de errores más frecuentes y partes de la oración afectadas por niveles del MCER recogidos en el corpus CORAF

Al igual que en la lista de frecuencias generales anteriores, para el error más predominante según las partes de la oración afectadas, creemos muy útil una comparación visual de la distribución de los errores en los distintos niveles estudiados.

Así las cosas, la distribución normalizada de los mismos puede observarse en el siguiente gráfico:

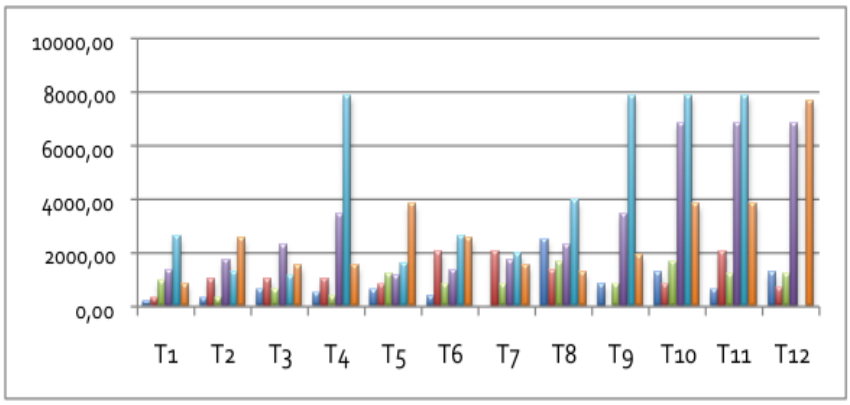

Gráfico 24: Distribución de los doce tipos de errores más frecuentes y partes de la oración afectadas por niveles del MCER.

ERROR	DESCRIPCIÓN
TIPO 1	Error léxico al formar un sustantivo de manera incorrecta por influencia de la L1.
TIPO 2	Error gramatical al seleccionar de forma incorrecta una preposición por hipergeneralización de reglas adquiridas.
TIPO 3	Error sintáctico al omitir un pronombre personal en la oración por interferencia de la L1.
TIPO 4	Error sintáctico al omitir un adverbio por interferencia de la L1.
TIPO 5	Error gramatical al seleccionar de forma incorrecta una preposición por interferencia de la L1
TIPO 6	Error léxico al seleccionar de forma inadecuada un sustantivo por una interferencia de la L1.
TIPO 7	Error ambiguo por utilización de la forma errónea del artículo indefinido por simplificación.

TIPO 8	Error sintáctico por alterar la posición del adverbio por la interferencia de la L1
TIPO 9	Error ambiguo en el uso del adjetivo por simplificación
TIPO 10	Error en la morfología flexiva de una forma verbal finita simple por simplificación.
TIPO 11	Error gramatical de causa ambigua y desconocida al seleccionar de forma errónea el artículo definido.
TIPO 12	Error gramatical por la utilización errónea del auxiliar en las formas verbales finitas compuestas, por hipergeneralización de reglas.

Tabla 27: Descripción de los doce errores más recurrentes que afectan a partes de la oración para todos los niveles del MCER.

De forma general, podemos observar que el número de errores decrece por niveles, aunque contemos con algunas alteraciones para el nivel C2 en ciertos tipos de error estudiados, como ya comentamos en la anterior clasificación general.

Los niveles que poseen más errores se ajustan generalmente, a los llamados de usuario básico (A1 y A2), y especialmente significativa es la posición del nivel B1, que suele ser donde podemos observar más errores en la gran mayoría de los tipos, y con una de las cifras de apariciones más altas, entre 13 y 15 errores. Tenemos que ser cuidadosos a la hora de interpretar estos errores, ya que el hecho de que sean cifras altas no depende sólo del grado de conocimiento de los aprendientes en el momento de la grabación, sino de sus propias características. En ocasiones, y dependiendo del tipo de error, es posible que nos encontremos con un aprendiente que produce reiteradamente el mismo error, lo que conlleva que la cifra aumente, pero no que esto sea representativo de su conjunto. Por consiguiente, y como venimos afirmando a lo largo del capítulo, en un futuro deberíamos de aumentar el número de muestras de aprendientes distintos para que los datos no estén ligados a especificidades de estos, y sean realmente extrapolables para el conjunto de los aprendientes de dicho nivel.

Así, el nivel más regular es el nivel C1, que obtiene en muchos de los tipos una cifra errores detectados bastante reducida, siendo incluso inexistente en alguno de ellos (como el tipo 12) y no superando la cifra máxima de siete. Lo que supone que pese a tener errores, parece que se está llevando a cabo un proceso de adquisición dentro de los cauces

esperados, ya que el error debe ir disminuyendo hasta la alcanzar el nivel de maestría.

El **nivel A1** tiene dificultades en los tipos 1, 6 y 11. Es cierto que el hecho de utilizar estructuras muy básicas puede que no muestre todos los errores que puedan cometer. Por tanto, ciñéndonos a lo analizado, sus errores más importantes se producen en el léxico (tipos 1 y 6), y generalmente, un mayor número en la producción de sustantivos con formas erróneas (tipo 1) que en la falsa selección de los mismos (tipo 6), por interferencia de la lengua materna, algo muy comprensible, al tener una competencia comunicativa muy limitada y restringida a los ámbitos que han estudiado en su contexto educativo.

Algo menor es la presencia de errores gramaticales de causa ambigua o desconocida por concordancia de género en los artículos definidos (tipo 11). Este tipo es normal que esté presente en los aprendientes de nivel más básico dado que es una de las características gramaticales que más se tarda en adquirir y que sigue planteando problemas en el resto de niveles, aunque muy a menudo, por causas distintas (como la simplificación).

De hecho, en estudios con corpus de aprendientes de lengua materna sueca, Bartning (2000) señala que la concordancia de género sigue planteando problemas en los niveles intermedio (*pre-advanced*) y avanzado. Además, puede demostrarse que el artículo definido se adquiere antes que el indefinido, puesto que el definido está marcado y establece diferenciaciones entre *le/la*. Y por otra parte, en aprendientes suecos se adquiere antes la forma masculina que la femenina.

En nuestro caso, los aprendientes de nuestro corpus, como bien podemos observar, siguen teniendo problemas en la concordancia del género, no sólo en artículos, sino también en adjetivos, en todos los niveles estudiados, aunque lógicamente se va reduciendo conforme se avanza en la adquisición. Como hemos comentado ya, esta concordancia suele ser uno de los puntos conflictivos en la interpretación de errores, ya que es una de las fuentes de errores ambiguos más usual.

El **nivel A2** refleja dificultades, como el A1, en la selección de sustantivos, los cuales construye de forma errónea (tipo 1), por una

interferencia de la lengua materna. Destaca en un menor número de ellos con problemas de falsa selección, relativos al significado (tipo 6).

También encuentra dificultades significativas en los tipos 10 y 12, donde es el nivel que más errores comete.

El tipo 10 se relaciona con una formación deficiente o errónea en la flexión de los verbos finitos simples por simplificación. Como hemos dicho anteriormente, la mayoría de los casos se relacionan con la conjugación del verbo '*étudier'*. Algunos ejemplos de ello son los siguientes:

*J'*étude* (A2M01)

*Je seulement *étude* (A2W02)

*Vxxx *étude* (A2M03)

El tipo de error 12 muestra una dificultad en la selección del auxiliar correcto en formas verbales compuestas por una tendencia a la generalización de las reglas. Es quizá, una de las características de los aprendientes de dicho nivel, que junto con los de B1, suponen más de la mitad de los errores cometidos (10 de un total de 14). Podemos dar cuenta de dicho aspecto en los siguientes ejemplos:

*C'est pour ça que *je m'ai marié avec lui* (A2W02)

*Et quand *j'ai allé* (A2W01)

**J'ai resté un petit peu* (A2W01)

El **nivel B1**, es con diferencia, el que más dificultades posee de las expuestas en la lista de errores frecuentes, siendo el que más número de estos aporta al conjunto.

Así, demuestra dificultades especialmente importantes en los tipos 2 (selección de la preposición de forma inadecuada por hipergeneralización), y 4 (omisión del adverbio por interferencia de la lengua materna). Es evidente que el sistema lingüístico que utilizan es cada vez más complejo, lo que produce ya la aparición de problemas en clases funcionales, y sobre todo, la emergencia de algunos procesos internos de reflexión sobre la lengua meta, que pueden interferir o provocar errores, como en el caso de las preposiciones. Sabemos que la

tendencia más extendida es utilizar la preposición que conocen en todos los contextos, sin importarles las restricciones del contexto o las diferenciaciones de su uso. Algo bastante usual en la preposición '*dans'* que utilizan en todos los contextos como 'en'. Ocurre, por ejemplo, en:

Il est dans l'école (B1W01), pretendiendo explicar así el nivel de escolarización de su hijo.

(j'ai fait) mes études supérieures dans la fac d'histoire en Cxxxx (B1W02)

**à jeudi, tu peux aller en boîte* (B1W03)

Además, mantiene dificultades en otros tipos como el 6 (falsa selección de un sustantivo por interferencia de la L1), el 7 (errores ambiguos por forma errónea en el uso de artículos indefinidos) y tipo 11 (concordancia de género con artículos definidos por razones desconocidas o ambiguas).

El **nivel B2**, suele ser, como vimos en la anterior clasificación general, uno de los más regulares. Mantiene ejemplos de errores en todos los tipos, pero sus cifras no son elevadas (no supera en ningún caso un número superior a 8, un 4,2 % de todos sus errores, y un valor normalizado de 857,2). Es el nivel que menos errores contiene de los tipos 3 (ausencia de pronombre personal por interferencia de la L1), 10 (flexión incorrecta de tiempos verbales finitos simples por simplificación) y 11 (concordancia de género en artículos definidos por causas ambiguas o desconocidas).

No obstante, sus mayores dificultades residen en la selección de preposiciones adecuadas por la interferencia de la lengua materna (tipo 5), aunque su número de apariciones tampoco es demasiado significativo (sólo 6, un 3,17 %, y un valor normalizado de 1.143), por lo que estaría dentro de las cifras normales del proceso de adquisición.

El **nivel C1**, como usuario competente, empieza ya a reflejar una mayor adecuación de su discurso, obteniendo cifras poco significativas de error en la mayoría de los tipos. Por comparación, es sin duda el nivel que mejores resultados obtiene, ya que en muchos de los tipos

encontramos tan solo un error, e incluso ninguno, como en el tipo 12 (uso del auxiliar adecuado en tiempos verbales compuestos).

Sus mayores dificultades, tampoco demasiado significativas, aparecen, como en el nivel B2, en el uso de las preposiciones adecuadas por la interferencia de la lengua materna. Algunos ejemplos que pueden mostrar estas dificultades son:

**de aller à la France* (C1M01)

De-Dès mon avis (C1M01)

Quelques cours par perfectionner (C1M02)

Ces concerts ont été patrocinés pour une bière (C1W03)

Finalmente, el **nivel C2** mantiene también un número de errores bastante reducido en todos los tipos descritos, lo que nos muestra que son errores que se siguen produciendo, porque van ligados a la producción de aprendientes muy concretos y de sus capacidades, pero que, al menos, se van reduciendo. No obstante, y al ser hablantes de nivel C2, deberían de contar aún con un número mucho menor de errores.

Según la lista de categorías y partes de la oración afectadas, las categorías en las que los aprendientes de nivel C2 encuentran más obstáculos tienen que ver, curiosamente, con dos errores por interferencia de la lengua materna. Uno de ellos es el tipo 1, que corresponde a un error léxico, y que es la categoría que se sitúa a la cabeza de las dificultades de prácticamente todos los niveles estudiados. Otro punto de dificultad es el tipo 8: error sintáctico por una alteración en la posición habitual del adverbio en la oración, como vemos en los ejemplos:

**Je n'ai retournée jamais* (C2W01)

Je trouve presque tout bien (C2M02)

Evidentemente, el desarrollo habitual de la adquisición traería aparejada una influencia de la lengua materna cada vez menor en los procesos de composición, formación o estructuración del enunciado en la lengua meta, pero como ya hemos mencionado anteriormente,

creemos en CORAF este hecho se debe, en parte, a las especificidades de los aprendientes que forman parte de nuestras entrevistas, que se apoyan demasiado en la traducción de elementos de la L1 para construir su discurso. Proceso, que por otra parte, es completamente razonable en el aprendiente, ya que se sirve de los conocimientos de su lengua materna y del filtro que supone para la construcción del nuevo conocimiento.

Por otra parte, como ya hemos señalado, existen muchas categorías de errores que aparecen ligadas a hablantes concretos, que producen el error de forma recurrente, lo que aumenta la frecuencia del mismo y su posición en las listas de errores más detectados.

4. Conclusiones

En el presente capítulo hemos presentado una aplicación basada en nuestro corpus oral: un análisis de errores de una cohorte de aprendientes hispanófonos de francés en contexto educativo. En el marco de dicho análisis, hemos mostrado el error desde tres planos distintos: lingüístico, descriptivo y etiológico. Además, hemos combinado los tres criterios citados para generar una lista de errores más frecuentes, que han sido descritos a nivel global, y a nivel específico para cada uno de los niveles del MCER. Finalmente, hemos puesto estos resultados en relación con la parte de la oración más comúnmente afectada por los errores.

Para el conjunto de nuestro corpus, hemos detectado aproximadamente 1.400 errores. En términos generales, y en relación con las muestras estudiadas, el análisis nos ha permitido llegar a las siguientes conclusiones:

- Desde el punto de vista lingüístico, los errores más frecuentes son de carácter gramatical, seguidos de los de carácter léxico y sintáctico. Las categorías más afectadas corresponden a los verbos, los artículos, las preposiciones y los sustantivos, siendo los menos frecuentes los errores en las conjunciones.
- Desde el punto de vista descriptivo, los errores más frecuentes afectan en primer lugar a las formas, y seguidamente aparecen problemas de selección incorrecta y de omisión de elementos lingüísticos requeridos por el enunciado en cuestión. Los

errores menos presentes son los de adición y los de colocación u orden incorrecto.

- Desde el punto de vista etiológico, se aprecia un porcentaje mínimo de errores de origen intralingüísticos en comparación con los de orden interlingüístico y los de tipo ambiguo. Los errores intralingüísticos se producen fundamentalmente por procesos de simplificación y generalización, y en menor medida por ignorancia de restricciones de las reglas, por analogía y por asociación cruzada. Los errores de tipo interlingüístico, como cabría esperar, se producen por la interferencia de la lengua materna, cuyo influjo se manifiesta, desde luego, en todos niveles del MCER, incluso en el nivel más avanzado como C2.
- Combinando todos los criterios anteriores en relación con la parte de la oración más afectada por los errores, obtenemos que los tres errores más frecuentes observados para todos los niveles corresponden a:
    - un error léxico por la formación errónea en sustantivos debido a una interferencia de la lengua materna;
    - un error gramatical por la inapropiada selección de la preposición por la puesta en práctica de un procedimiento intralingüístico de hipergeneralización;
    - y un error sintáctico por la omisión del pronombre personal por la influencia de la lengua materna.

En términos generales, hemos constatado que el nivel más regular en la comisión de errores es el B2, que también es el nivel con menor número de errores detectados. Entre los que aportan más errores al conjunto destaca el nivel C1, que suele estar a la cabeza en muchos de los tipos analizados.

4. CONCLUSIONES GENERALES

El presente trabajo versa sobre un análisis de errores de aprendientes de Francés como Lengua Extranjera (FLE) en contexto educativo, basado en un corpus oral creado en el marco del propio estudio: CORAF (*Corpus Oral de Aprendientes de Francés*). Para alcanzar los objetivos del proyecto, hemos procedido como sigue:

1. Hemos desarrollado sucesivamente las tareas de recopilación, transcripción y alineación del corpus;

2. Hemos aplicado al corpus una modalidad de análisis de errores de corte tradicional, utilizando tres criterios básicos de estudio: un criterio lingüístico (que describe el tipo de error según la parte de la oración afectada y el nivel de estudio de la lengua); un criterio descriptivo (que muestra la distancia entre la forma errónea y el enunciado habitual para un hablante nativo), y un criterio etiológico (que indaga las causas más probables de aparición del error).

El corpus que hemos generado se compone de 30 entrevistas, distribuidas entre los seis niveles de dominio de la lengua que establece el MCER (*Marco Común Europeo de Referencia* para la enseñanza de las lenguas*)*. Las grabaciones tienen una duración que supera las siete horas, y reúnen un total de 61.092 palabras, 33.915 de las cuales corresponden a los aprendientes. El corpus ha sido transcrito siguiendo las convenciones del LLI-UAM (*Laboratorio de Lingüística Informática de la Universidad Autónoma de Madrid*), las cuales han sido enriquecidas con aportes de otros corpus de aprendientes, que también fueron creados para el marcado de fenómenos propios de la interlengua o lengua del aprendiente. El resultado es un corpus digital que sigue los estándares habituales para este tipo de recursos, resulta fácilmente consultable y explotable mediante las herramientas propias del campo de la Lingüística de Corpus.

Nuestra primera caracterización de la lengua del aprendiente de FLE por medio de un análisis de errores de las muestras recogidas en CORAF nos ha proporcionado un total de 1.400 errores aproximadamente, que se distribuyen y categorizan de forma variada,

dependiendo de los niveles de dominio del aprendiente y de los criterios utilizados en cada caso para el análisis.

Nuestra convicción de la utilidad de nuestra contribución para los avances de la investigación y de la educación no nos impide ser conscientes de sus limitaciones que, a nuestro entender, se centran en los siguientes aspectos:

- El tamaño de nuestro corpus oral es bastante reducido, y esto limita las posibilidades de extrapolar sus conclusiones a todos los aprendientes de FLE de lengua materna española. Para dar un valor más general a nuestros resultados, hubiera sido necesario contar con un número más elevado de palabras y con una cohorte más amplia de consultantes.

- Nuestro estudio también ganaría en fiabilidad con un sistema más elaborado de determinación del nivel de los aprendientes en el momento en el que se producen las grabaciones. Por un lado, se podría haber realizado un pre-test que estableciera el nivel real de los aprendientes. Por otro, también podrían haberse realizado las grabaciones simultáneamente en todas las Escuelas Oficiales de Idiomas, de forma que todos los aprendientes se encontrasen a la misma altura del curso académico, aunque, en realidad, no creemos que unos meses más o menos de estudio tengan una incidencia significativa sobre el nivel de dominio de una L2.

- Podría haberse extendido el recuento de errores según el total de categorías producidas, lo que nos hubiera dado una visión más acertada de la predominancia de un error en una determinada categoría. Es posible que aunque se produzcan muchos errores, estos no sean de un tipo frecuente, conocimiento que sin duda, puede cambiar nuestra percepción sobre el mismo.

- En las conclusiones que presentamos, no todo es objetivo, pues al menos en lo que se refiere a la atribución de la causa del error, el análisis está mediatizado por la interpretación personal por parte de la investigadora. Tal parecer, pese a apoyarse en el conocimiento y la experiencia, tiene un carácter inevitablemente subjetivo. Esto quiere decir que los resultados del presente trabajo ganarían a ser contrastados con las opiniones de

otros investigadores. Dicha discusión podría producirse ya sometiendo los resultados al análisis de otros investigadores, ya comparando las conclusiones con las de otros estudios que utilicen asimismo el análisis de errores automático o asistido por ordenador (CEA).

Respecto a las taxonomías utilizadas para el análisis de errores, queremos resaltar que sería interesante revisar en profundidad la composición de las taxonomías disponibles en el mercado, con el fin de lograr que delimiten mucho más el criterio lingüístico, ya que el hecho de que un número significativo de errores sea susceptibles de incluirse en más de una categoría, y también de que muchas de las categorías interfieran entre sí, son inconvenientes que limitan la claridad de los resultados.

También creemos necesario que se revise el criterio etiológico, pues al ser tan específico, genera tipologías de errores muy concretos, lo que redunda en un número muy bajo de apariciones en todos los niveles de dominio.

Una vez señaladas, entendemos que de manera objetiva, las limitaciones del trabajo, consideramos, no obstante, que supone una contribución relevante a los estudios de carácter docente o científico sobre la interlengua de FLE, por cuatro motivos fundamentalmente:

- Contribuimos a la mejora del análisis de la interlengua con un corpus oral de aprendientes hispanófonos de FLE, cuando no existe en la actualidad ningún otro estudio parecido para el citado público específico.
- El corpus CORAF ofrece un nuevo campo de explotación de datos, abriendo la vía a otros estudios con objetivos y enfoques analíticos diferentes;
- Los resultados de nuestra investigación pueden convertirse en la base de estudios comparativos con otros recursos de la misma categoría para aprendientes de FLE con lenguas maternas distintas al español.
- Puede ser utilizado por parte de investigadores y docentes, tanto como herramienta de trabajo de uso directo con

aprendientes, como para la generación de material pedagógico adaptado a las dificultades particulares del público objeto de nuestro estudio.

En general, pensamos que nuestro trabajo puede contribuir de manera decisiva a la mejora de los métodos actuales de enseñanza de FLE para hispanófonos, permitiendo el desarrollo de unas prácticas mucho más específicas y adaptadas a las necesidades de los aprendientes.

Todo lo anterior nos invita a reflexionar sobre el potencial de prolongación de nuestra investigación a nivel personal, perspectiva que inicialmente desearíamos que se centrara en dos ejes fundamentales:

- Ampliación del corpus con una mayor cohorte de sujetos consultantes para cada uno de los niveles expuestos, lo que a su vez nos permitiría completar nuestro análisis de errores;
- Realización de nuevos estudios de la interlengua introduciendo nuevas perspectivas, a saber, el análisis del discurso y el análisis de la actuación. Dichos estudios podrían complementarse con propuestas pedagógicas destinadas a futuros docentes, para la mejora de la competencia comunicativa de los aprendientes.

BIBLIOGRAFÍA

ABOUDA, L. et BAUDE, O. (2005): "Constituer et exploiter un grand corpus oral: choix et enjeux théoriques. Le cas de ESLO". En: *Actes du colloque international d'Albi Langages et Signification (CALS): « Corpus en lettres et sciences sociales : des documents numériques à l'interprétation »*, juillet 2006, Albi [En línea]: <http://www.revue-texto.net/Parutions/Livres-E/Albi-2006/Actes_ALBI-0> [Consulta: 12/06/2008]

AIJMER. K. (Ed.) (2009): *Corpora and Language Teaching*. Amsterdam: John Benjamins.

ALTET, M. (1997): *Les pédagogies de l'apprentissage*. París: PUF.

ÁLVAREZ GONZÁLEZ, S. y MARTÍNEZ GARCÍA, J. A. (2007): "La evolución de la enseñanza y aprendizaje de lenguas extranjeras (francés) en la senda de las Nuevas Tecnologías". En: *Didáctica (Lengua y Literatura)*, volumen 19, pp.47-74. [En línea:] <http://www.ucm.es/BUCM/revistas/edu/11300531/articulos/DIDA0707110047A.PDF> [Consulta: 22/10/2008].

ANTONIADIS, G. et PONTON, C. (2004): "MIRTO: un système au service de l'enseignement des langues". En: *Cinquième colloque des Usages des Nouvelles Technologies pour l'Enseignement des Langues Etrangères (UNTELE 2004)*. 16-20 mars 2004. Compiègne [En línea]: <http://w3.u-grenoble3.fr/lidilem/labo/file/UNTELE.pdf > [Consulta: 08/11/2008].

ANTONIADIS, G.; FAIRON, C.; GRANGER, S.; MEDORI, J. et ZAMPA, V. (2006): "Quelles machines pour enseigner la langue?". En: MARTENS, C., FAIRON, C., WATRIN, P. (eds.) : *TALN06 : Verbum ex machina actes de la 13e conférence sur le Traitement Automatique des Langues Naturelles*. Vol. 2, pp. 793-805 [en línea]: <http://www.noe-kaleidoscope.org/group/idill/repository/Antoniadis.pdf> [Consulta: 05/11/2008].

ANTONIADIS, G., PONTON, C. et ZAMPA, V. (2007): "De la nécessité du TAL dans les EIAH en langues : Les cas EXXELANT et MIRTO". En: *Actes EIAH 2007*. 27-29 juin 2007, Lausanne (Suisse), [en

línea]: <http://w3.u-grenoble3.fr/ponton/perso/docs/EIAH07.pdf > [Consulta: 05/11/2008].

ASTON. G. (1997): "Small and large corpora in language learning". En: Lewandowska-Tomaszczyk, B. and P.J. Melia (eds.): *PALC'97: practical applications in language corpora*, Lodz: Lodz University Press.

ASTON, G. (2001a): *Learning with Corpora*. Houston: Athelstan.

ASTON, G. (2001b): "Text categories and corpus users: A response to David Lee". En: *Language Learning & Technology*, vol. 5, nº 3. [En línea]: <http://www.sslmit.unibo.it/~guy/astonrelee.htm> [Consulta: 23/08/2009].

ASTON, G. (2002): "The learner as corpus designer". En: Kettemann, B. and Marko, G. (eds.) (2002): *Teaching and Learning by doing Corpus Analysis*. Amsterdam: Rodopi. pp. 9-25

ASTON, G., BERNARDINI, S. And STEWART, D. (Eds.) (2004): *Corpora and Language Learners*. Amsterdam/Philadelphia: John Benjamins.

BARR, D., LEAKEY, J. and RANCHOUX, A. (2005): "Told like it is! An evaluation of an integrated oral development pilot project". En: *Language Learning & Technology*, September 2005, volume 9, Number 3, pp. 55-78. [En línea]: <http://llt.msu.edu/vol9num3/barr/> [Consulta: 04/07/2008].

BARTNING, I (2000): "Gender agreement in L2 French: pre-advanced vs. advanced learners". En: *Studia Linguistica*, 54 (2), pp. 225-237. [En línea]: <http://onlinelibrary.wiley.com/doi/10.1111/1467-9582.00062/full> [Consulta: 23/11/2009].

BARTNING, I. et SCHLYTER, S. (2004): "Itinéraires acquisitionnels et stades de développement en français L2", en: *French Language Studies* 14, pp. 281-299. [En línea]: <http://www.foreignpolicybulletinmonitor.com/action/displayAbstract?fromPage=online&aid=276786&fulltextType=RA&fileId=S0959269504001802> [Consulta: 23/06/2009].

BAUDE, O. (coord.) (2006): *Corpus Oraux. Guide de Bonnes Pratiques*. Orléans: Presses Universitaires d'Orléans-CNRS Éditions.

BEAUGRANDE, R. de (2000): "Large corpora and applied linguistics. H.G. Widowsson versus J. Sinclair". En BATTANER, M.P. y LÓPEZ, C. (Eds.): *VI Jornada de Corpus Linguistics: Corpus Linguistics i ensenyament de llengües*. Barcelona: IULA, Universitat Pompeu Fabra, pp. 87-104.

BEECHING, K. (1997): "French for Specific Purposes: The case for Spoken Corpora". En: *Applied Linguistics*, volume 18, num. 3, pp. 374-391. [En línea]: <http://applij.oxfordjournals.org/content/18/3/374.short > [23/09/2009].

BERNARDINI, S. (2003): "Designing a Corpus for Translation and Language Teaching: the CEXI Experience", en *TESOL QUARTERLY*, *37 (*3), pp. 528-537. [En línea]: <http://www.jstor.org/pss/3588403> [Consulta: 15/10/2009].

BIBEAU, R. (2005): "Les TIC à l'école: proposition de taxonomie et analyse des obstacles à leur intégration". En: *Le matériel didactique et pédagogique : soutien à l'appropriation ou déterminant de l'intervention éducative*. Québec: Les Presses de l'Université Laval. Pp. 297-325. [En línea]: <http://www.robertbibeau.ca/conference.html > [Consulta: 15/06/2008].

BIBER, D. (1993a): "Using Register-Diversified Corpora For General Language Studies". En: *Computacional Linguistics*, Volume 19, Issue 2, pp. 219-241. [En línea]: <http://portal.acm.org/citation.cfm?id=972472> [Consulta: 15/04/2009].

BIBER, D. (1993b): "Representativeness in Corpus Design", en *Literary and Linguistic Computing,* Vol.8, nº4, pp. 243-257. [En línea]: <http://llc.oxfordjournals.org/cgi/content/short/8/4/243> [Consulta: 12/09/2009].

BIBER, D. (2004): "Conversation Text-Types: A multi-dimensional analysis" En: *Actes des 7e Journées Internationales d'Analyse statistique des Dones Textuelles, JADT 2004*. Louvain: Presses Universitaires de Louvain. [En línea]: <http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2004/pdf/JADT_000.pdf > [Consulta: 15/04/2009].

BIBER, D. and CONRAD, S. (2001): "Quantitative Corpus-Based Research: Much More Than Bean Counting". En: *Tesol Quarterly*, vol.35, nº2, pp.331-336.

BLANCHE-BENVENISTE, C. (1997): "De l'utilité du corpus linguistique", en *Revue française de linguistique appliquée*, Vol. I, 2, pp. 25-42.

BLANCHE-BENVENISTE, C. (1998) : *Le français parlé, études grammaticales*. Paris : CNRS.

BLANCHE-BENVENISTE, C. et BILGER, M. (1999): "Français parlé-oral spontané. Quelques réflexions". En: *Revue Française de Linguistique Appliquée*, Dossier *"L'oral spontané"*, vol. IV-2/déc. (pp. 21-30). [En línea]: <http://icar.univ-lyon2.fr/ecole_thematique/contaci/documents/bilger_cappeau/CBB-Bilger.pdf> [Consulta: 12/01/2008].

BLANCHE-BENVENISTE, C. (2002): "Compréhension multilingue et connaissance de sa propre langue". [En línea]. < http://ancilla.unice.fr/~brunet/pub/claire.html> [Consulta: 12/12/2007].

BLANCHE-BENVENISTE, C. (2003): *Approches de la langue parlée en français*. Paris: Ophrys.

BLOCK, D. and CAMERON, D. (2002): *Globalization and Language Teaching*. London: Routledge.

BOURIGAULT, D. & FABRE, C. (2000): "Approche linguistique pour l'analyse syntaxique de corpus". En: *Cahiers de Grammaire, « Sémantique et Corpus »,* pp.131-151. [En línea:] <http://w3.erss.univ-tlse2.fr:8080/index.jsp?perso=bourigault&subURL=bourigault-pub.html> [Consulta: 16/10/2008].

BOULTON, A. and WILHELM, S. (2006): "Habeant Corpus-they should have the body. Tools learners have the right to use", en: *Proceedings of 27th GERAS CONFERENCE: Teaching and Corpora*. Lorient, France: Université de Bretagne-Sud, 23-25 March 2006. [En línea]: <http://hal.archives-ouvertes.fr/docs/00/11/42/77/PDF/2006_GERAS_corpus.pdf> [Consulta: 21/09/2008].

BOULTON, A. (2007a): "Esprit de Corpus: Promouvoir l'exploitation de corpus en apprentissage des langues". En: *Actes des Journées de la Linguistique de Corpus 2007,* pp. 37-46. [En línea]: <http://en.scientificcommons.org/55079081> [Consulta: 12/09/2008].

BOULTON, A. (2007b): "DDL Is in the Details… and in the Big Themes". En: M. Davies, P. Rayson, S. Hunston & P. Danielsson (eds.): *Proceedings of the Corpus Linguistics Conference: CL2007*. [En línea]: <http://www.corpus.bham.ac.uk/corplingproceedings07/paper/126_Paper.pdf> [Consulta: 23/09/2008].

BOULTON, A. (2008a): "Looking for empirical evidence of data-driven learning at lower levels." En: Lewandowska-Tomaszczyk, B. (dir.): *Corpus Linguistics, Computer Tools, and Applications – State of the Art*. Frankfort: Peter Lang, 581-598.

BOULTON, A. (2008b): "But where's the proof? The need for empirical evidence for data-driven learning." En: Edwards, M. (dir.): *Proceedings of the BAAL Annual Conference 2007*. London : Scitsiugnil Press. [En línea]: <http://hal.inria.fr/docs/00/32/67/04/PDF/2007_boulton_BAAL_proof.pdf> [Consulta: 26/11/2011].

BOULTON, A. (2009a): "Data-Driven learning: On Paper, In Practice". En: HARRIS, T. And MORENO JAÉN, M. (eds.): *Corpora in Language Teaching*. Berna: Peter Lang (Linguistic Insights).

BOULTON, A. (2009b): "Data-driven Learning: Reasonable Fears and Rational Reassurance". En: *Indian Journal of Applied Linguistics*, volume 35, nº1, pp. 81-106. [En línea]: <http://hal.archives-ouvertes.fr/hal-00326990/ > [Consulta: 23/09/2010].

BOULTON, A. (2009c): "Testing the limits of data-driven learning: language proficiency and training." En: *ReCALL* 21/1 : 37-51. [En línea]: <http://davies-linguistics.byu.edu/ling485/for_class/teaching/boulton_limits.pdf > [Consulta: 21/10/2011].

BOULTON, A. (2009d): "Corpora for all? Learning styles and data-driven learning." En: *5th Corpus Linguistics Conference Proceedings*. Liverpool, Royaume-Uni : Universidad de Liverpool, 20-23 juillet.

BOULTON, A. (2010a): "Consultation de corpus et styles d'apprentissage", en: *Cahiers de l'APLIUT*, 29/1. Pre-print version. [En línea]: <http://hal.archives-ouvertes.fr/hal-00448993/> [Consulta: 26/03/2010].

BOULTON, A. (2010b): "Data-driven learning: Taking the computer out of equation". En: *Language Learning*, 60, 3, pp. 534-572. [En línea]: <http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9922.2010.00566.x/full> [Consulta: 25/02/2011].

BOULTON, A. y TYNE, H. (2008): "Learning with corpora: changing learning practices." En: *4th Inter-Varietal Applied Corpus Studies (IVACS) Group Conference: Applying Corpus Linguistics*. Limerick, Ireland: Universidad de Limerick, 13-14 juin.

BRAUN, S. (2005): "From pedagogically relevant corpora to authentic language learning contents". En: RECALL 17 (1), pp. 47-64. [En línea]: < http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=305550> [Consulta: 25/09/2010].

BRAUN, S. (2006): "ELISA – A pedagogically enriched corpus for language learning purposes". En: S. Braun, K. Kohn & J. Mukherjee (Eds.): *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Frankfurt/M: Lang, pp.25-47.

BRAUN, S. (2007): "Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora", en: *ReCALL* 19 (3), pp. 307-328. [En línea]: < http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=1313948> [Consulta: 12/11/2010].

BROWN, A. (2009): "Students' and teachers' perceptions of effective foreign language teaching: a comparison of ideals." En: *Modern Language Journal* 93/1, pp. 46-60. [En línea]: <http://onlinelibrary.wiley.com/doi/10.1111/j.1540-4781.2009.00827.x/full> [Consulta: 27/11/2011].

BROWN, H. D. (2007): *Principles of Language Learning and Teaching.* (5th edition) New York: Pearson Education.

CAMPILLOS LLANOS, L.; GOZALO GÓMEZ, P. y MORENO SANDOVAL, A. (2007): "El corpus C-ORAL-ROM en la enseñanza de ELE". En: BALMASEDA MAESTU, E. (ed.): *Actas del XVII congreso internacional de ASELE: Las destrezas orales en la enseñanza del español.* Logroño. 27-30 de septiembre de 2006. Logroño: Servicio de publicaciones de la Universidad de La Rioja.

CAPPEAU, P. et SEIJIDO, M. (2005) : *Les corpus oraux en français (inventaire 2005, v.1.0)*, [en línea]. Délégation Générale à la Langue Française et aux Langues de France. <http://www.culture.gouv.fr/culture/dglf/recherche/corpus_parole/Presentation_Inventaire.pdf> [Consulta: 15/10/2007].

CARTER, R. And McCARTHY, M. (1994): "Grammar and the spoken language". En: *Proceedings of the Annual Meeting of the Teachers of English to Speakers of Other Languages*. Baltimore, March 1994.

CASSIDY, S. (2004). "Learning styles: an overview of theories, models and measures." En: *Educational Psychology* 24/4, pp. 419-444. [En línea]: <http://www.acdowd-designs.com/sfsu_860_11/LS_OverView.pdf > [Consulta: 26/10/2010].

CAWS, C. (2009): "Contexte et culture en enseignement du FLS: De la création d'un corpus à son exploitation didactique", en: *Mélanges CRAPEL*, nº 31, pp. 206-220. [En línea]: <http://revues.univ-nancy2.fr/melangesCrapel/article_melange.php3?id_article=330> [Consulta: 23/11/2009].

CENTRE FOR ENGLISH CORPUS LINGUISTICS (Universidad de Lovaina) [En línea]: <http://www.uclouvain.be/en-cecl.html > [Consulta: 12/06/2011].

CENTRO VIRTUAL INSTITUTO CERVANTES (2001): *Marco de referencia europeo para el aprendizaje, la enseñanza y la evaluación de lenguas*. Recurso electrónico. [En línea]: <http://cvc.cervantes.es/ensenanza/biblioteca_ele/marco/> [Consulta: 21/10/2011].

CERMÁK, F. (2002): "Today's corpus linguistics: Some open questions". En: *International Journal of Corpus Linguistics, 7* (2), 265-282.

CHAMBERS, A. (2005): "Integrating corpus consultation in language studies". En: *Language Learning & Technology*. May 2005, volume 9, number 2, pp. 111-125 [en línea]: <http://llt.msu.edu/vol9num2/chambers/> [Consulta: 21/10/2008].

CHAMBERS, A. (2009): "Les corpus oraux en français langue étrangère: authenticité et pédagogie". En: *Mélanges CRAPEL*, numéro 31, pp. 16-33, [en línea]: < http://revues.univ-nancy2.fr/melangesCrapel/article_melange.php3?id_article=319> [Consulta: 12/12/2009].

CHAMBERS, A., CONACHER, J. and LITTLEMORE, J. (dir.) (2004): *ICT and Language Learning : Integrating Pedagogy and Practice*. Birmingham: Universidad de Birmingham Press.

CHAN, P.-T. y LIOU, H.C. (2005): "Effects of web-based concordancing instruction on EFL students' learning of verb–noun collocations." En: *Computer Assisted Language Learning* 18/3, pp. 231-251. [En línea]: <http://taylorandfrancis.metapress.com/link.asp?target=contribution&id=X787L574514W87M5> [Consulta: 20/10/2010].

CHANIER, T. et CIEKANSKI, M. (2010): "Utilité du partage des corpus pour l'analyse des interactions en ligne en situation d'apprentissage: un exemple d'approche méthodologique autour d'une base de corpus d'apprentissage". En: *Revue ALSIC (Apprentissage des Langues et des Systèmes d'Information et de Communication),* volume 13. [En línea]: <http://alsic.revues.org/index1666.html> [Consulta: 12/03/2011].

CHAPELLE, C. A. (1997): "CALL in the year 2000: still in search of research paradigms?" En: *Language Learning & Technology*, vol. 1 Nº 1, July 1997, pp.19-43. [En línea]: <http://llt.msu.edu/vol1num1/chapelle/default.html> [Consulta: 12/10/2008]

CHAPELLE, C. A. (2001): *Computer Applications in Second Language Acquisition. Foundations for teaching, testing and research*. Cambridge: Cambridge University Press.

CHEVRIER, J., FORTIN, G. Y OTROS (2000): « Le style d'apprentissage » En: *Education et francophonie, vol XXVIII :1. Revista de la Association Canadienne d'Éducation de Langue Française*. Universidad de Ottawa.

CONRAD, S. (1996): "Investigating Academic Texts With Corpus-Based Techniques: An example from Biology". En: *Linguistics and Education*, vol. 8, pp. 299-326.

CONSEIL DE L'EUROPE (2001): *Un cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*. [En línea]. Paris : Didier. <http://www.coe.int/T/DG4/Portfolio/documents/cadrecommun.pdf >

[Consulta: 04/11/2007].

CORDER, S. PIT (1967): "The signifiance of learner's errors", en: *IRAL*, 5, 4.

CORDER, S. PIT (1981): *Error Analysis and Interlanguage*. First edition. Oxford: Oxford University Press.

COSMIDES, L. y TOOBY, J. (1992): "Cognitive adaptations for social exchange." En: Barkow, H., L. Cosmides y J. Tooby (dir.): *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford : Oxford University Press, pp. 163-228.

COURTILLON, J. (2003): *Élaborer un cours de FLE*. Paris: Hachette Français Langue Étrangère.

CRAPEL: *Centre de Recherche d'Applications Pédagogiques En Langues*. Université de Nancy 2 (France). <http://www.univ-nancy2.fr/CRAPEL/> [Consulta: 14/02/2010].

CRESTI, E., MONEGLIA, M., BACELAR, F., SANDOVAL, A. M., VÉRONIS, J., MARTIN, P., CHOUCRI, K., MAPELLI, V., FALAVIGNA, D., and CID, A. (2002). "The C-ORAL-ROM project: New methods for spoken language archives in a multilingual romance corpus". En: M.C. RODRÍGUEZ and C. SUÁREZ ARAUJO (Eds.): *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2002)* (pp. 2-10). Paris: ELRA.

CRESTI, E. y MONEGLIA, M. (eds.) (2005): *C-ORAL-ROM Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam: John Benjamins.

CRESTI, E. and SCARANNO, A. (2000): "Sur la notion de parlé spontané", en: M. Bilger (éd.), *Corpus. Méthodologie et applications linguistiques*. Paris: Champion.

CRESSWELL, A. (2007): "Getting to 'know' connectors? Evaluating data-driven learning in a writing skills course." En: Hidalgo, E., L. Quereda y J. Santana (dir.): *Corpora in the Foreign Language Classroom*. Amsterdam: Rodopi, pp. 267-287.

DAGNEAUX, E., DENNESS, S. and GRANGER, S. (1998): "Computer-aided error analysis". En: *System,* volume 26, issue 2 (June 1998), pp. 163-174. [En línea]: <http://www.sciencedirect.com/science?_ob=PublicationURL&_tockey=%23TOC%235955%231998%23999739997%2319261%23FLP%23&_cdi=5955&_pubType=J&view=c&_auth=y&_acct=C000050221&_version=1&_urlVersion=0&_userid=10&md5=5c38bc5f92cf25c243dca006930de0e1> [Consulta: 21/10/2010].

DAVIES, G. (2002): "CALL (Computer assisted language learning)", en: *Subject Centre for Languages, Linguistics and Area Studies Good Practice Guide*. [En línea]: <http://www.llas.ac.uk/resources/gpg/61> [Consulta: 25/11/2010].

DEBAISIEUX, J. M. (1997): "Pour une approche micro et macro du français parlé dans la formation des enseignants de FLE". En: *Mélanges CRAPEL*, nº 23, pp. 27-53, [en línea]:

<http://revues.univ-nancy2.fr/melangesCrapel/article_melange.php3?id_article=67>
[Consulta: 20/02/2008].

DEBAISIEUX, J.M. et REGENT, O. (1999): "Un outil multimedia pour apprendre à apprendre les langues étrangères". En: *Mélanges CRAPEL*, nº 24, pp. 45-58, [en línea]:

<http://revues.univ-nancy2.fr/melangesCrapel/article_melange.php3?id_article=37>
[Consulta: 24/10/2008].

DEBAISIEUX, J.M. (2005) : "Les corpus oraux : Situation, exploitation linguistique, bilan et perspectives". En: *De la linguistique de corpus à la relation « partie/tout »*, Scolia, n°19, Publications de l'Université Marc Bloch – Strasbourg 2, pp. 9-40, [en línea]: <http://halshs.archives-ouvertes.fr/docs/00/14/91/41/PDF/3._scolia.pdf> [Consulta: 22/05/2008].

DEBAISIEUX, J.M (2009): "Des documents authentiques oraux aux corpus: un défi pour la didactique du FLE". En: *Mélanges CRAPEL*, numéro 31, pp. 36-56, [en línea]: <http://revues.univ-nancy2.fr/melangesCrapel/article_melange.php3?id_article=320> [Consulta: 16/12/2009].

DEBROCK, M., FLAMENT-BOISTRANCOURT, D. et GEVAERT, R. (1999): "Le manque de 'naturel' des interactions verbales du non-francophone en français. Analyse de quelques aspects à partir du corpus LANCOM". En: *Faits de Langues*, volume 7, num. 13, pp. 46-56. [En línea]: <http://www.persee.fr/web/revues/home/prescript/article/flang_1244-5460_1999_num_7_13_1237 > [Consulta: 23/07/2009].

DEGACHE, C. (2004): "Interactions asynchrones et appropriation dans un environnement d'apprentissage collaboratif des langues (Galanet) " En: BAQUÉ, L. et TOST, M. A. (eds.): *Repères et applications (IV)*, Universitat Autònoma de Barcelona [en línea]: <http://www.galanet.be/publication/fichiers/dc2004.pdf > [Consulta: 12/05/08].

DEGACHE, C. (2005): "Comprendre la langue de l'autre et se faire comprendre ou la recherche d'une alternative communicative: le projet Galanet". En: S. Borg et M. Drissi (éds.): *Approches pédagogiques et instruments didactiques pour le plurilinguisme*, *Synergies-Italie*, n°2-2005, Programme mondial de diffusion scientifique francophone en réseau, Gerflint, [en línea]: <http://ute2.umh.ac.be/galanet/publication/fichiers/dc2005.pdf> [Consulta: 25/04/2008].

DEGACHE, C. (2006): "Entorno multimedia, autoformación y enseñanza de lenguas". En: *Estudios de Lingüística del Español (ELiEs)*,

volumen 24. Traducción de Paloma Garrido, [en línea]: <http://elies.rediris.es/elies24/degache.htm> [Consulta: 20/12/2007].

DÉLÉGATION GÉNÉRALE À LA LANGUE FRANÇAISE ET AUX LANGUES DE France (2006): "Corpus de la parole", en : *Langues et cité*, nº 6, mayo 2006, Ministère de la culture et de la communication. [En línea] <http://www.culture.gouv.fr/culture/dglf/Langues_et_cite/Langues_cite6.pdf> [Consulta: 21/11/2007].

DEMAIZIÈRE, F. (2007): "Didactique des langues et TIC: Les aides à l'apprentissage". En: *Revue ALSIC (Apprentissage des Langues et des Systèmes d'Information et de Communication)*, volume 10, nº1, pp. 5-21. [En línea]: <http://alsic.revues.org/index220.html> [Consulta: 03/08/2010].

DESMET, P. and HÉROUGUEL, A. (2005): "Les enjeux de la création d'un environnement d'apprentissage électronique axé sur la compréhension orale à l'aide du système auteur IDIOMA-TIC". En: *Revue ALSIC (Apprentissage des Langues et des Systèmes d'Information et de Communication)*, volume 8, 2005, pp. 281-303. [En línea]: <http://alsic.u-strasbg.fr/v08/desmet/alsic_v08_12-poi4.htm> [Consulta: 03/12/2007].

DIAZ NEGRILLO, A., FERNÁNDEZ-DOMÍNGUEZ, J. (2006): "Error-tagging systems for learner corpora". En: *RESLA*, 19, pp. 83-102. [En línea]: < http://www.mendeley.com/research/error-tagging-systems-for-learner-corpora/>. [Consulta: 21/06/2010].

DOUGLAS BROWN, H. (2007): *Principles Of Language Learning and Teaching*. (5th Edition) New York: Pearson Education.

DOYLE, W. & RUTHERFORD, B. (1984): "Classroom research on matching learning and styles and teaching styles". En: *Theory into practice,* 23, 1, pp. 20-25.

DUDA, R. y RILEY, P. (dir.). (1990): *Learning Styles*. Nancy : Presses Universitaires de Nancy.

DULAY, H., BURT, M. and KRASHEN, S. (1982): *Language Two*. New York: Oxford University Press.

DURKHEIM, E. (1922): *Éducation et sociologie*. Paris: Les Presses universitaires de France, Collection "Le sociologue".

EHRMAN, M. (2008): "Personality and good language learners." En: Griffiths, C. (dir.): *Lessons from Good Language Learners*. Cambridge: Cambridge University Press, pp. 61-72.

EHRMAN, M., B. LEAVER Y R. OXFORD (dir.). (2003): "Individual Differences: Advancing Knowledge". En: *System,* volume 31/3, pp. 313-330. [En línea]: <http://www.sciencedirect.com/science/journal/0346251X/31/3> [Consulta: 20/10/2010].

ELLIS, N. (2002): "Frequency effects in language processing: a review with implications for theories of implicit and explicit language acquisition". En: *Studies in Second Language Acquisition*, 24 (2), pp. 143-188. [En línea]: <http://www.lotschool.nl/files/schools/archief/Winterschool%20Nijmegen%202007/dabrowska/Ellis%202002.pdf > [Consulta: 21/03/2010].

ELLIS, N. C. (2008): "The Dynamics of Second Language Emergence: Cycles of Language Use, Language Change, and Language Acquisition". En: *The Modern Language Journal*, 92, pp. 232-249. [En línea]: <http://onlinelibrary.wiley.com/doi/10.1111/j.1540-4781.2008.00716.x/full> [Consulta: 21/05/2010].

ELLIS, R. (1997): *Second Language Acquisition*. (17th edition) Oxford: Oxford University Press.

ELLIS, R. and BARKHUIZEN, G. (2005): *Analysing Learner Language*. Oxford: Oxford University Press.

ÉQUIPE DELIC (2004): "Présentation du *Corpus de référence du français parlé*". En: *Recherches sur le français parlé*, 18, pp. 11-42. [En línea]: <http://sites.univ-provence.fr/~veronis/pdf/2004-presentation-crfp.pdf > [Consulta: 21/02/2008].

ESTLING VANNESTÅL, M. y H. LINDQUIST. (2007): "Learning English grammar with a corpus: experimenting with concordancing in a university grammar course." En: *ReCALL*, 19/3 , pp. 329-350. [En línea]: <http://journals.cambridge.org/action/displayAbstract;jsessionid=8AD

3355A4FEE61A09C2E658099BFF49B.journals?fromPage=online&aid=1313956 > [Consulta: 20/11/2010].

FAERCH, C. and KASPER, G. (eds.) (1984): *Strategies in Interlanguage communication*. London: Longman.

FARR, F. (2008): "Evaluating the Use of Corpus-based Instruction in a Language Teacher Education Context: Perspectives from the Users". En: *Language Awareness*, volume 17, nº1, pp.25-43. [En línea]: <http://www.informaworld.com/smpp/content~db=all~content=a907041267> [Consulta: 21/09/2010].

FELDER, R. (1993): "Reaching the second tier: learning and teaching styles in college science education." En: *Journal of College Science Teaching*, 23/5, pp. 286-290. [En línea]: <http://www4.ncsu.edu/unity/lockers/users/f/felder/public/Papers/Secondtier.html > [Consulta: 19/06/2009].

FELDER, R. Y HENRIQUES, E. (1995): "Learning and teaching styles in foreign and second language education." En: *Foreign Language Annals,* 28/1, pp. 21-31. [En línea]: <http://www4.ncsu.edu/unity/lockers/users/f/felder/public/Papers/FLAnnals.pdf > [Consulta: 19/06/2009].

FELDER, R. Y L. SILVERMAN. (1988). "Learning and teaching styles in engineering education." En: *Engineering Education*, 78/7, pp. 674-681. [En línea]:

<http://www4.ncsu.edu/unity/lockers/users/f/felder/public/Papers/LS-1988.pdf > [Consulta: 19/06/2009].

FELDER, R. Y SPURLIN, J. (2005): "Applications, reliability, and validity of the Index of Learning Styles." En: *International Journal of Engineering Education*, 21/1, pp. 103-112. [En línea]: <http://www4.ncsu.edu/unity/lockers/users/f/felder/public/ILSdir/ILS_Validation(IJEE).pdf> [Consulta: 19/06/2009].

FERNÁNDEZ LÓPEZ, S. (1997). *Interlengua y análisis de errores en el aprendizaje del español como lengua extranjera.* Madrid: Edelsa.

FLLOC: *French Learner Language Oral Corpora*. Universidad de Southampton (England). <http://www.flloc.soton.ac.uk/> [Consulta: 12/01/2010].

FLOWERDEW, L. (1998): "Integrating 'Expert' and 'Interlanguage' Computer Corpora Findings on Causality: Discoveries for Teachers and Students". En: *English for Specific Purposes*, volume 17, nº4, pp. 329-345. [En línea]: <http://www.sciencedirect.com/science/article/pii/S0889490697000148> [Consulta: 23/04/2008].

FLOWERDEW, L. (2005): "An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: countering criticisms against corpus-based methodologies". En: *English for Specific Purposes*, 24, pp. 321-332. [En línea]: <http://www.renevenegas.cl/mexico/bibliografia/Flowerdew%20%282005%29%20ESP.pdf> [Consulta: 23/04/2008].

FLOWERDEW, L. (2008): "Pedagogic value of corpora: a critical evaluation." En: Frankenberg-Garcia, A. (dir.): *Proceedings of the 8th Teaching and Language Corpora Conference*. Lisboa : Associação de Estudos e de Investigação Científica do ISLA-Lisboa, pp. 115-119.

FRANKENBERG-GARCIA, A. (2005): "A peek into what today's language learners as researchers actually do." En: *International Journal of Lexicography*, 18/3, pp. 335-355. [En línea]: <http://ijl.oxfordjournals.org/content/18/3/335.short> [Consulta: 23/04/2008].

FRANKENBERG-GARCIA, A., FLOWERDEW, L. And ASTON G. (Eds.) (2011): *New Trends in corpora and Language Learning*. London: Continuum.

GABRIELATOS, C. (2005): "Corpus and Language Teaching: Just a fling or wedding bells?" En: *Teaching Language as a Second or Foreign Language* (TESL-EJ), vol. 8, nº 4, A-1, [en línea]: <http://www.tesl-ej.org/ej32/a1.html> [Consulta: 22/10/2008].

GASS, S. M. and SELINKER, L. (2001): *Second Language Acquisition: an introductory course* (2nd edition). Mahwah, New Jersey: Lawrence Erlbaum Associates.

GAVIOLI, L. (2000): "Some thoughts on the Problem of Representing ESP through small corpora". En: KETTEMANN, Bernhard and Georg MARKO (Eds.): *Proceedings of the Fourth International Conference on Teaching and Language Corpora*, Graz 19-24 July, 2000, pp. 293-303(11). [En línea]: <http://www.ingentaconnect.com/content/rodopi/lang/2002/00000042/00000001/art00021> [Consulta: 15/09/2009].

GAVIOLI, L. and ASTON, G. (2001): "Enriching reality: language corpora in language pedagogy". En: *ELT Journal*, volume 55/3, pp. 238-246. [En línea]: <http://203.72.145.166/ELT/files/55-3-3.pdf> [Consulta 15/09/2009].

GERMAIN, A. and MARTIN, P. (2000): "Présentation d'un logiciel de visualisation pour l'apprentissage de l'oral en langue seconde". En: *Revue Alsic (Apprentissage des Langues et des Systèmes d'Information et de Communication)*, volume 3, numéro 1, juin 2000, pp. 61-76. [En línea] : <http://toiltheque.org/Alsic_volume_1-7/Num5/germain/alsic_n05-rec7.htm> [Consulta : 12/06/2009].

GERMAIN, C. (1993): *Évolution de l'enseignement des langues: 5000 ans d'histoire*. Paris: Clé International.

GODWIN-JONES, B. (2001): "Emerging technologies: Tools and Trends in Corpora Use for Teaching and Learning". En: *Language Learning & Technology*, vol. 5, num. 3, pp. 7-12, [en línea]: <http://llt.msu.edu/vol5num3/emerging/> [Consulta: 03/12/2008].

GONZÁLEZ, A., DE LA MADRID, G., ALCÁNTARA, M., DE LA TORRE, R., MORENO, A. (2004): "Orality and difficulties in the transcription of a spoken corpus". En: *Proceedings of the Workshop on Compiling and Processing Spoken Corpora*. LREC-2004, Lisboa.

GRANFELDT, J., NUGUES, P. (2007): "Évaluation des stades de développement en français langue étrangère". En: *Actes du colloque TALN 2007*, Toulouse, 5-8 Juin, pp. 357-366. [En línea]: <http://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=539356&fileOId=625980> [Consulta: 18/04/2010].

GRANGER, S., HUNG, J. and PETCH-TYSON, S. (2002): *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins.

GRANGER, S. (2003a): "Error-tagged Learner Corpora and CALL: A Promising Synergy", en: *CALICO Journal*, 20 (3), pp. 465-480. [En línea]: <https://www.calico.org/memberBrowse.php?action=article&id=289> [Consulta: 25/09/2010].

GRANGER. S. (2003b): "The corpus approach: a common way forward for Contrastive Linguistics and Translation Studies". En: Granger Sylviane, Lerot Jacques, Petch-Tyson Stephanie: *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*, Amsterdam & Atlanta: Rodopi.

GRANGER, S. (2004): "Computer Learner Corpus Research: Current Status and Future Prospects". En: *Language and Computers, Applied Corpus Linguistics. A Multidimensional Perspective*, pp. 123-145(23). [En línea]: <http://www.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/Downloads/Indianapolis%20status%20&%20prospects.pdf> [Consulta: 24/02/2010].

GRANGER, S. (2009): "The contribution of learner corpora to second language acquisition and foreign language teaching. A critical evaluation". En: AIJMER. K. (Ed.) (2009): *Corpora and Language Teaching*. Amsterdam: John Benjamins, pp. 13- .

GREMMO, M.J. and HOLEC, H. (1990): "La compréhension orale: un processus, un comportement". En: *Le Français dans le Monde*, L'approche cognitive, février-mars 1990. Paris, [en línea]: <http://www.epc.univ-nancy2.fr/EPCT_F/pdf/La%20compOrale.pdf > [Consulta: 22/04/2008].

GREVISSE, M. (1993): *Le bon usage*. Paris: Duculot. 13ª edición.

GRISHMAN, R. (1986): *Computational Linguistics: An introduction*. Cambridge: Cambridge University Press.

GUERNIER, M.C. et SAUTOT, J.P. (2004): "Celui qui ne parle pas apprend-il aussi?". En: *Actes du colloque international « Faut-il parler pour apprendre ? »*, Mars 2004, Arras, pp.1-8., [en línea]: <http://w3.u-grenoble3.fr/lidilem/labo/file/guernier_mariececile.pdf> [Consulta: 09/01/2008].

GUILQUIN, G., GRANGER, S. and PAQUOT, M. (2007): "Learner corpora: the missing link in EAP pedagogy". En: *Journal of English for*

Academic Purposes, volume 6 (4), pp. 319-335. [En línea]: <http://www.mendeley.com/research/learner-corpora-the-missing-link-in-eap-pedagogy/> [Consulta: 21/09/2009].

GUIRAO, J. M., MORENO, A., GONZÁLEZ, A., DE LA MADRID, G. and ALCÁNTARA, M. (2004): "Relating linguistic units to socio-contextual information in a spontaneous speech corpus of Spanish". En: *Corpus Linguistics Across the World*. Amsterdam: Rodopi.

GRIES, S. Th. and STEFANOWITSCH, A. (Eds.) (2009): *Corpus Linguistics and Linguistic Theory,* 5 (1). Berlin : Mouton de Gruyter.

HAMEL, M. J. et MILICEVIC, J. (2007): "Analyse d'erreurs lexicales d'apprenants du FLS: Démarche empirique pour l'élaboration d'un dictionnaire d'apprentissage". En: *Canadian Journal of Applied Linguistics (CJAL)/Revue canadienne de linguistique appliquée (RCLA)*, Vol 10, No 1. pp.25-45. [En línea]: <http://ojs.vre.upei.ca/index.php/cjal/article/viewArticle/256>. [Consulta: 21/09/2010].

HEGELHEIMER, V., TOWER, D. (2004): "Using CALL in the classroom: Analyzing students' interactions in an authentic classroom". En: *System 32*, Issue 2, pp. 185-205. [En línea]: <http://www.sciencedirect.com/science/article/pii/S0346251X04000211> [Consulta: 20/04/2009].

HOLEC, H. (1970): "Compréhension orale en langue étrangère". En: *Mélanges Pédagogiques.* Nancy: CRAPEL, Université de Nancy 2. [En línea]: <http://revues.univ-nancy2.fr/melangesCrapel/IMG/pdf/holec.pdf> [Consulta: 22/10/2008].

HOLEC, H. (1990a): "Des documents authentiques, pour quoi faire?". En: *Mélanges Pédagogiques,* Nancy : CRAPEL, Université de Nancy 2, [en línea]: <http://revues.univ-nancy2.fr/melangesCrapel/IMG/pdf/5holec-2.pdf> [Consulta: 09/01/2008].

HOLEC, H. (1990b): "Qu'est-ce qu'apprendre à apprendre?". En *Mélanges Pédagogiques,* Nancy : CRAPEL, Université de Nancy 2. [En línea]: <http://revues.univ-nancy2.fr/melangesCrapel/IMG/pdf/6holec-3.pdf> [Consulta: 09/01/2008].

HOUSSAYE, J. (2000): *Le triangle pédagogique. Théorie et pratiques de l'éducation scolaire*. Berna: Peter Lang. (3ª ed. , 1ª ed. 1988).

HUNSTON, S. (2002): *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

HYMES, D. (1972): "On communicative competence". En: J. B. Pride and J. Holmes (eds.): Sociolinguistics. Harmondsworth, England: Penguin Books.

IDE, N. and VÉRONIS, J. (Eds.). (1995): *The Text Encoding Initiative: Background and Context*. Dordrecht: Kluwer Academic.

IZUMI, E., UCHIMOTO, K. and ISAHARA, H (2004): "SST speech corpus of Japanese learners' English and automatic detection of learners errors", en: *ICAME Journal*, nº 28, pp. [En línea]: <http://icame.uib.no/ij28/index.html> [Consulta: 25/09/2010].

JAMES, C. (1998): *Errors in Language Learning and Use. Exploring Error Analysis*. London: Longman.

JOHANSSON, S. (1995): "Mens Sana in Corpore Sano: On the role of corpora in linguistic research". *The European Messenger, IV* (2), 19-25.

JOHANSSON, S. (2009): "Some thoughts on corpora and second-language acquisition." En: Aijmer, K. (dir.): *Corpora and Language Teaching*. Amsterdam: John Benjamins, pp. 33-44.

JOHNS, T. (1991): "Should you be persuaded- Two samples of data-driven learning materials". En: *ELR Journal*, volume 4, pp.1-16.

KASZUBSKI, P. (2008): "A guided collaboration tool for online concordancing with EFL EAP learners." En: Frankenberg-Garcia, A. (dir.): *Proceedings of the 8th Teaching and Language Corpora Conference*. Lisboa : Associação de Estudos e de Investigação Cientifíca do ISLA-Lisboa, pp. 167-175.

KEFFE, J.W. (1987): *Learning Style theory and practice*. Reston, Virginia: National Association of Secondary School Principals.

KIRSCHNER, P., SWELLER, J. and CLARK, R. (2006): "Why minimal guidance during instruction does not work : an analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based

teaching." En: *Educational Psychologist*, 41/2 , pp. 75-86. [En línea]: <http://www.tandfonline.com/doi/abs/10.1207/s15326985ep4102_1> [Consulta: 27/10/2010].

KOLB, D.A. (1984): *Experiential Learning. Experience as the Source of Learning and Development*. Englewood Cliffs, New Jersey: Prentice-Hall.

KRAIF, O. (2004): "Propositions pour l'intégration à la didactique des langues d'outils issus du traitement automatique des langues". En: DEGACHE, C. (dir.): *Intercompréhension en langues romanes, LIDIL*, N° 28, Université Stendhal, Grenoble, [en línea]: <http://w3.u-grenoble3.fr/lidilem/labo/file/Kraif2004Lidil28.PDF> [Consulta: 05/12/2007].

KRAIF, O. (2006): "Qu'attendre de l'alignement de corpus multilingues?". En: *Revue Traduire*, 4e Journée de la traduction professionnelle. Société Française des Traducteurs, numéro 210, pp.17-37. [En línea]: <http://w3.u-grenoble3.fr/lidilem/labo/file/RevuetraduireV1.pdf> [Consulta: 21/01/2008].

L'HAIRE, S. et VANDEVENTER FALTIN, A. (2003): "Diagnostic d'erreurs dans le projet FreeText". En: *Revue ALSIC- Apprentissage des Langes et Systèmes d'Information et de Communication*, volume 6, numéro 2, pp. 21-37. [En línea] : <http://alsic.revues.org/index2219.html> [Consulta: 24/02/2010].

LAKSHAMANAN, U. and SELINKER, L. (2001): "Analysing interlanguage: how do we know what learners know?". En: *Second Language Research,* 17, issue 4, pp. 393-420. [En línea]: <http://slr.sagepub.com/content/17/4/393.short> [Consulta: 23/03/2010].

LANDURE, C. Et BOULTON. A (2010): "Corpus et autocorrection pour l'apprentissage des langues", en *Revue ASp*, nº 57. Pre-print version. [En línea]: <http://hal.archives-ouvertes.fr/hal-00448970/> [Consulta: 26/03/2010].

LARSEN-FREEMAN, D. and LONG, M. H. (1991): *An introduction to second language acquisition research*. Harlow (England): Longman.

LAVID, J. (2005): *Lenguaje y Nuevas Tecnologías*. Madrid: Cátedra.

LEECH, G. (2006): "New resources or Just Better Old Ones? The Holy Grail of Representativeness". En: Marianne Hundt, Nadja Nesselhauf and Carolin Biewer (eds.): *Language and Computers, Corpus Linguistics and the Web.*, pp. 133-149. Amsterdam: Rodopi.

LEGENDRE, R. (1988): *Dictionnaire actuel de l'éducation*. París/Montreal: Larousse.

LEHUEN, J., LEMEUNIER, T. et LUZZATI, D. (2002): "Acquisition et étude d'un corpus FLE. Vers une analyse automatique des erreurs". En: *Actes du Quatrième colloque des Usages des Nouvelles Technologies dans l'Enseignement des Langues Etrangères (UNTELE 2002)*, 28-30 mars, Compiègne, [en línea]: <http://www-lium.univ-lemans.fr/~lemeunie/publications/Lehuen-et-al-UNTELE-2002.pdf> [Consulta: 28/03/2008].

LEWIS, J. (2006): *Connecting Corpora to Learner Style : To what Extent is the Effectiveness of an Online Corpus-Based Approach to Grammar Learning Dependent on whether Students Prefer to Learn Grammar Deductively or Inductively ?* Proyecto fin de máster. Porto : Universidad de Porto. [En línea] <http://www.fc.up.pt/fcup/contactos/teses/t_020370029.pdf> [Consulta: 19-06-2009].

LIGHTBOWN, P. M. (2003): "SLA research in the classroom/SLA research for the classroom". En: *Language Learning Journal*, nº 28, pp. 4-13.

LIGHTBOWN, P. M. and SPADA, N. (2006): *How Languages are Learned.* Oxford: Oxford University Press.

LITZINGER, T., S. LEE, J. WISE and R. FELDER. (2007): "A psychometric study of the Index of Learning Styles." En: *Journal of Engineering Education*, 96/4 , pp. 309-319. [En línea]: <https://eng.kuleuven.be/onderwijs/onderwijsondersteuning/1988_felder.pdf > [Consulta: 19/06/2009].

MACKEY, A. and GASS, S. M. (2005): *Second Language Research: Methodology and Design.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

MANGENOT, F. (1998): "Classification des apports d'Internet à l'apprentissage des langues". En: *Revue Alsic (Apprentissage des Langues et*

des Systèmes d'Information et de Communication), volume 1, numéro 2, décembre 1998, pp.133-146, [en línea]: <http://alsic.u-strasbg.fr/Num2/mangenot/alsic_n02-pra1.htm> [Consulta: 25/02/2008].

MANGENOT, F. (2000): "L'intégration des TICE dans une perspective systémique". En: *Les Langues Modernes* (novembre 2000), "Les nouveaux dispositifs d'apprentissage". Paris, Association des Professeurs de Langues Vivantes. [En línea]: <http://w3.u-grenoble3.fr/espace_pedagogique/publicat.htm> [Consulta: 05/05/2008].

MANGENOT, F. (2005): « Quelles compétences, quelles formations, quels métiers liés aux TICE ? ». En: LES CAHIERS DE L'ASDIFLE n°16, LES METIERS DU FLE. Paris, Association de didactique du français langue étrangère, p. 163-176. [En línea]: <http://w3.u-grenoble3.fr/espace_pedagogique/publicat.htm> [Consulta: 05/05/2008].

MARCOS MARÍN, F. (1994): *Informática y Humanidades*. Madrid: Gredos.

MAYER, R. (2004): "Should there be a three-strike rule against pure discovery learning ? The case for guided methods of instruction." En: *American Psychologist,* 59/1, pp. 14-19.

McCARTHY, M. (2006): *Explorations in Corpus Linguistics*. New York: Cambridge University Press.

McCARTHY, M. (2008): "Accessing and interpreting corpus information in the teacher education context". En: *Language Teaching*, 41:4, 563-574. [En línea]: <http://journals.cambridge.org/abstract_S0261444808005247> [Consulta: 21/04/2009].

McCARTHY, M. and CARTER, R. (2001): "Size Isn't Everything: Spoken English, Corpus, and the Classroom". En: *TESOL Quarterly*, volume 35, nº1, pp. 337-340. [En línea]: <http://www.jstor.org/pss/3587654> [Consulta: 12/05/2009].

McENERY, T. and WILSON, A. (1996): *Corpus Linguistics*. Edinburgh: Edinburgh University Press. [En línea]:

<http://www.ling.lancs.ac.uk/monkey/ihe/linguistics/contents.htm> [Consulta: 28/11/2007].

McENERY, T. and WILSON, A. (1997): "Teaching and Language Corpora", en: *ReCALL*, Volume 9, nº1, May 1997, pp. 5-12. [En línea]: <http://www.eurocall-languages.org/recall/pdf/rvol9no1.pdf#page=5> [Consulta: 12/19/2008].

McENERY, T. and XIAO, R. (2010): "What corpora can offer in language teaching and learning". En: *Handbook of Research in Second Language Teaching and Learning*, London and New York: Routledge.

MISHAN, F. (2004): "Authenticating corpora for language learning: a problem and its resolution". En: *ELT Journal*, volume 58/3, pp. 219-227. [En línea]: <http://eltj.oxfordjournals.org/content/58/3/219.abstract> [Consulta: 25/02/2009].

MITCHELL, R. and MYLES, F. (2004): *Second Language Learning Theories* (2nd edition). London: Hodder Arnold.

MORENO SANDOVAL, A. (1998): *Lingüística Computacional. Introducción a los modelos simbólicos, estadísticos y biológicos*. Madrid: Síntesis.

MORENO SANDOVAL, A. (2002): "La evolución de los corpus de habla espontánea: la experiencia del LLI-UAM". En: *Actas de las Segundas Jornadas de Tecnologías del Habla*, Universidad de Granada.

MORENO SANDOVAL, A. (2003): "Los corpus orales del LLI-UAM: Primera generación y segunda generación". En: *La Musa Digital*, número 3. Especial: *Proceedings of the Computers Literature and Philology Conference (CLIP 2002).*

MORENO SANDOVAL, A. y URRESTI, J. (2005): "El proyecto C-ORAL-ROM y su aplicación a la enseñanza del español". En: *ORALIA*, vol. 8, pp. 81-104. Madrid: Arco Libros.

MUÑOZ, C. (Ed.) (2000): *Segundas Lenguas. Adquisición en el aula.* Barcelona: Ariel.

MYLES, F. (2004a): "Second Language Acquisition (SLA) research: its signifiance for learning and teaching issues". En: *Subject Centre for Languages, Linguistics and Area Studies web.* University of Southampton. [en

línea]: <http://www.llas.ac.uk/resources/gpg/421> [Consulta: 12/05/2008].

MYLES, F. (2004b): "French second language acquisition research: setting the scene". En: *French Language Studies,* 14, pp. 211-232. [En línea]: <http://journals.cambridge.org/action/displayAbstract;jsessionid=DAA61FB8D28E8D07B5445BDB68F6A367.journals?fromPage=online&aid=276780> [Consulta: 12/05/2008].

MYLES, F. (2005): "Interlanguage corpora and second language acquisition research". En: *Second Language Research* 21, 4, pp.373-391. [En línea]: <http://www.corpus4u.org/upload/forum/20051127210207 65.pdf> [Consulta: 11/05/2009].

NEL, C. (2008): "Learning style and good language learners." En: Griffiths, C. (dir.): *Lessons from Good Language Learners*. Cambridge : Cambridge University Press, pp. 49-60.

NICHOLLS, D. (2004): "The Cambridge Learner Corpus – error coding and analysis for lexicography and ELT", en: *Proceedings of Corpus Linguistics, 2003*, pp. 572-582. [En línea]: <http://ucrel.lancs.ac.uk/publications/CL2003/contents.htm> [Consulta: 25/09/2010].

NICOLÁS, C. (2003): "Una propuesta de utilización de corpus orales en la enseñanza de segundas lenguas". En: *La Musa Digital*, nº 3. Especial: *Proceedings of the Computers Literature and Philology Conference (CLIP 2002).*

O'KEEFFE, A. and FARR, F. (2003): "Using language corpora in Initial Teacher Education: Pedagogic Issues and Practical Applications". En: *Tesol Quarterly*, vol. 37, nº 3, pp. 389-418. [En línea]: <http://www.ingentaconnect.com/content/tesol/tq/2003/00000037/00000003/art00002> [Consulta: 21/10/2009].

O'KEEFFE, A. and McCARTHY, M, (Ed.) (2010): *The Routledge Handbook of Corpus Linguistics*. London: Routledge.

O'KEEFFE, A., McCARTHY, M. and CARTER, R. (2007): *From Corpus to Classroom. Language use and language teaching.* Cambridge: Cambridge University Press.

OSBORNE, J. (2004): "Top-Down and Bottom-Up Approaches to Corpora in Language Teaching", en: *Language and Computers. Applied Corpus Linguistics. A Multidimensional Perspective*. Edited by Ulla Connor and Thomas A. Upton, pp. 251-265(15). [En línea]: <http://www.ingentaconnect.com/content/rodopi/lang/2004/00000052/00000001/art00015> [Consulta: 15/11/2009].

OXFORD, R.L. and ANDERSON, N.J. (1995): "A crosscultural view of learning styles". En: *Language Teaching*, 28, pp. 201-215. [En línea]: <http://journals.cambridge.org/action/displayAbstract;jsessionid=A694A7F05ED3A8A15233469241C25E00.journals?fromPage=online&aid=2584512> [Consulta: 23/10/2010].

POISSON-QUINTON, S., MIMRAM, R., MAHÉO-LE COADIC, M. (2004): *Grammaire expliquée du français*. Paris: Clé International.

PORQUIER, R. (1977): "L'analyse d'erreurs: Problèmes et perspectives". En : *Études de Linguistique Appliquée*, numéro 25, janvier-mars 1977, pp. 25-43.

POTHIER, M. (1997): "Hypermédia et autonomie", en OUDART, P. (coord..): *Multimédia, réseaux et formation,* Le Français Dans le Monde, Recherches et Applications, EDICEF, pp. 85-93.

PRAVEC, N. A. (2002): "Survey of learner corpora". En: *ICAME Journal* 26, pp. 81-114. [En línea]: <http://icame.uib.no/ij26/pravec.pdf> [Consulta: 11/12/2009].

PY, B. (1984): "L'analyse contrastive: histoire et situation actuelle". En: *Le Français Dans le Monde*, nº 238, pp. 44-52.

RASTIER, F. (2004) : « Enjeux épistémologiques de la linguistique de corpus », en: *Actas de las IIª Jornadas de Lingüística de Corpus de Lorient*. Presse Universitaires de Rennes. Publicada en : *Texto !,* juin 2004. Rubrique Dits et inédits. [En línea] : <http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html> [Consulta : 21/10/2010].

REID, J. (dir.). (1995): *Understanding Learning Styles in the Second Language Classroom*. Londres : Prentice Hall.

REINERT, H. (1976): "One picture is worth a thousand words? Not necessarily!" En: *The Modern Language Journal*, volume 60, pp. 160-168.

[En línea]: <http://education.jhu.edu/newhorizons/strategies/topics/Learning%20Styles/picture.html > [Consulta: 19/06/2009].

RICHARDS, J. (1971a): "A non-contrastive approach to error analysis". En: *English Language Teaching* 25, 3. [En línea]: <http://eric.ed.gov/PDFS/ED037721.pdf> [Consulta: 23/10/2010].

RICHARDS, J. (1971b): "Error Analysis and Second Language Strategies". En: *Language Science*, 17. [En línea]: <http://eric.ed.gov/PDFS/ED048579.pdf> [Consulta: 23/10/2010].

RICHARDS, J.C. (Ed.) (1974): *Error Analysis. Perspectives on Second Language Acquisition*. London: Longman.

RICHARDS, J. C. and RODGERS, T.S. (2001): *Approaches and Methods in Language Teaching*. (2nd edition). Cambridge: Cambridge University Press.

RIDING, R. and RAYNER, S. (1998): *Cognitive Styles and Learning Strategies: Understanding Style Differences in Learning and Behaviour*. London: David Fulton Publishers.

RIVENC, P. (2005): "Espoirs et limites de l'étude quantitative des corpus (dialogues et vocabulaires dits « disponibles »)". En: *Actes du Colloque International du Français Fondamental, Corpus Oraux, Contenus d'Enseignement, 50 ans de travaux et enjeux*. Lyon, [en línea]: <http://colloqueff.ens-lsh.fr/pdf/Rivenc_Paul-2.pdf> [Consulta: 28/06/2008].

RÖMER, U. (2006): "Pedagogical Applications of Corpora: Some Reflections on the Current Scope and a Wish List for Future Developments", en: *ZAA* 54.2, pp. 121-134. [En línea]: <http://uteroemer.com/ZAA%202006%20Ute%20Roemer.pdf> [Consulta: 12/02/2010].

ROSEN, É. (2005): "La mort annoncée des « quatre compétences » – pour une prise en compte du répertoire communicatif des apprenants en classe de FLE". En: *Glottopol, Revue de sociolinguistique en ligne*, nº6, pp. 120-133. [En línea]: <http://www.univ-rouen.fr/dyalang/glottopol/numero_6.html> [Consulta: 20/10/2010].

RULE, S., MARSDEN, E., MYLES, F., MITCHELL, R. (2003): "Constructing a database of French interlanguage oral corpora", en:

Archer, D., Rayson, P., Wilson, E. & McEnery, T. (eds.): *Proceedings of the Corpus Linguistics 2003 Conference*, UCREL Technical Papers no. 16, pp. 669-77, University of Lancaster.

SÁNCHEZ LOBATO, J. y SANTOS GARGALLO, I. (Dir.) (2004): *Vademécum para la formación de profesores: enseñar español como segunda lengua (L2) – lengua extranjera (LE)*. Alcobendas (Madrid): Sociedad General Española de Librería.

SANTOS GARGALLO, I. (1993): *Análisis Contrastivo, Análisis de Errores e Interlengua en el marco de la Lingüística Contrastiva*. Madrid: Síntesis.

SANZ GIL, M. (2003a): *Las Tecnologías de la Información y de la Comunicación y la autonomía de aprendizaje de lenguas. Análisis crítico y estudio de casos en el aprendizaje del FLE*. Tesis doctoral. [En línea]: <http://www.tdx.cbuc.es/TESIS_UJI/AVAILABLE/TDX-0628104-113234//msanz.pdf> [Consulta: 20/10/2007].

SANZ GIL, M. (2003b): "Aprender francés a través de la red. Los retos del aprendizaje en autonomía.". En: *El texto como encrucijada: estudios franceses y francófonos* / coord. por Ignacio Iñarrea Las Heras, María Jesús Salinero Cascante, Vol. 2, 2003, pp. 717-730. [En línea]: <http://dialnet.unirioja.es/servlet/articulo?codigo=1011643> [Consulta: 26/11/2008].

SANZ MIGUEL, C. (1999): *El Libro de las Preposiciones. Diccionario de dificultades de uso de las preposiciones en el idioma francés*. Toledo: Azacanes.

SALABERRI, R. (2001): "The use of Technology for Second Language Learning and Teaching: A Retrospective". En: *The Modern Language Journal*, 85, I, 2001, pp. 39-56. [En línea]: <https://webspace.utexas.edu/mrs2429/www/Salaberry2001MLJ.pdf> [Consulta: 21/10/2010].

SCHACHTER, J. and CELCE-MURCIA, M. (1977): "Some Reservations Concerning Error Analysis", en: TESOL Quarterly, vol. 11, nº 4, pp. 441-451. [En línea]: <http://www.jstor.org/pss/3585740> [Consulta: 12/09/2010].

SCHMECK, R. (1983): "Learning styles of college students". En: R. F. Dillon and R. Schmeck (eds.): *Individual differences in cognition, vol. 1*. New York: Academic Press, pp. 233-279.

SELINKER, L. (1972): "Interlanguage". En: *International Review of Applied Linguistics in Language Teaching*, 10, 3, pp. 209-231. [Traducción al español: "La interlengua", en Muñoz Liceras, J. (comp.) (1991): *La adquisición de las lenguas extranjeras*. Madrid: Visor, pp. 79-101)].

SILBERTZTEIN, M. et TUTIN, A. (2005): "NooJ: un outil TAL de corpus pour l'enseignement des langues. Application pour l'étude de la morphologie lexicale en FLE". En: *Revue ALSIC (Apprentissage des Langues et des Systèmes d'Information et de Communication)*, volume 8, 2005, pp.123-134. [En línea]: <http://alsic.u-strasbg.fr/v08/silberztein/alsic_v08_20-rec11.htm> [Consulta: 06/11/2008].

SINCLAIR, J. (1991): *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

SINCLAIR, J. M. (1997): "Corpus linguistics at the millennium". En: J. Kohn, B. Rüschoff, & D. WOLFF (Eds.): *New Horizons in CALL. Proceedings of European Association for Computer Assisted Language Learning (EUROCALL 96)*, pp. 1-10. Szombathely: Bersenyi Daniel College.

SOLOMAN, B. and R. FELDER (1996): *Index of Learning Styles Questionnaire*. [En línea]: <http://www.engr.ncsu.edu/learningstyles/ilsweb.html> [Consulta: 19/06/2009].

SOUBRIÉ, T. (2005): "Le présentiel allégé à l'université pour les grands groupes : un dispositif au service de l'autonomisation des apprenants". En : ACTES DU COLLOQUE SIF 2005, Paris, 12-13 décembre 2005. [En línea]: <http://w3.u-grenoble3.fr/lidilem/labo/file/SIF05.pdf> [Consulta: 21/05/2008].

SOUBRIÉ, T. (2006): "Dispositif de formation ouverte à l'utilisation des TIC dans l'enseignement/apprentissage du FLE en Master", En: *Actes du 7e colloque européen sur l'Autoformation : faciliter les apprentissages autonomes*, Auzeville (France), mai 2006.

STOCKER, C. (1921): "French speech-tunes and the phonograph", en: *The Modern Language Journal*, 5, pp. 267-270.

STORTI, G. (2001): *Comparaison de la méthode situationnelle et de l'approche communicative en didactique de langues*. Edición digital.

TAN, M. (2005): "Authentic language or language errors? Lessons from a learner corpus". En: *ELT Journal*, volume 59 (2), pp.126-134. [En línea]: <http://eltj.oxfordjournals.org/content/59/2/126.short > [Consulta: 10/06/2009].

TANIMURA, M., TAKEUCHI, K. and ISAHARA, H. (2005): "From Learners' Corpora to Expert Knowledge Description: Analyzing Prepositions in the NICT JLE (Japanese Learner English) Corpus", en: *Proceedings of IWLet, 2004: An Interactive Workshop on Language e-Learning*, pp. 139-147. [En línea]: <http://en.scientificcommons.org/893402> [Consulta: 25/09/2010].

TARDIF, J. (1996): "Une condition incontournable aux promesses des NTIC en apprentissage: une pédagogie rigoureuse", Conférence d'ouverture du 14ème colloque de l'AQUOPS, *Une pédagogie sans frontières, AQUOPS*. [En línea]: <http://www.ac-grenoble.fr/occe26/printemps/tardif/pedagogie.htm> [Consulta: 12/06/2009].

TONO, Y. (2003): "Learner corpora: design, development and applications*"*. En: *Proceedings of Corpus Linguistics 2003 Conference (CL 2003),* Lancaster, England. [En línea]: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=?doi=10.1.1.115.6849> [Consulta: 10/10/2009].

TOMÉ, M (2006): "L'enseignant de FLE et les ressources Internet", en: Revista *Çedille*, revista de estudios franceses, nº2 (2006), pp. 114-133. [En línea]: <http://redalyc.uaemex.mx/pdf/808/80800208.pdf > [Consulta: 12/06/2010]

TOMÉ. M. (2009a): "Enseignement des langues, communication et compétences orales sur le web actuel". En: Revista *Çedille*, revista de estudios franceses, nº5 (abril de 2009), pp.347-370. [En línea]: <http://redalyc.uaemex.mx/src/inicio/ArtPdfRed.jsp?iCve=80811192019> [Consulta: 12/06/2010]

TOMÉ, M. (2009b): "Productions orales, *weblogs* et projet de télécollaboration avec le web 2.0. pour l'enseignement du français (FLE)". En: Revue *ALSIC (Apprentissage des Langues et des Systèmes*

d'Information et de Communication), volume 12, 2009, pp. 90-108. [En línea]: <http://alsic.revues.org/index1279.html> [Consulta: 22/10/2010].

TORDERA ILLESCAS, J. C. (2010): *Lingüística Computacional y Anáfora.* Valencia: Servicio de Publicaciones de la Universidad de Valencia.

TRANSANA, [en línea] <http://www.transana.org>. [Consulta: 06/11/2007].

TURNBULL, J. y BURSTON, J. (1998): "Towards independent concordance work for students : lessons from a case study." En*: ON-CALL,* 12/2, pp. 10-21.

TYNE, H. (2009): "Corpus oraux par et pour l'apprenant", en: *Mélanges CRAPEL*, nº31, pp. 92-111. [En línea]: <http://revues.univ-nancy2.fr/melangesCrapel/article_melange.php3?id_article=323> [Consulta: 13/12/2009].

UNESCO (1984): *Tesaurus de la Educación*. París: UNESCO/OIE.

UNESCO (2004): *Las tecnologías de la información y la comunicación en la formación docente*. División de Educación Superior. UNESCO. [En línea]: <http://unesdoc.unesco.org/images/0012/001295/129533s.pdf > [Consulta: 18/05/2010].

VÁZQUEZ, G. (1991): *Análisis de errores y aprendizaje de español lengua extranjera. Análisis, explicación y terapia de errores transitorios y fosilizables en el proceso de aprendizaje de español como lengua extranjera en cursos universitarios para hablantes nativos de alemán.* Berlin: Peter Lang.

VÉRONIQUE, D. (2008): "À l'intersection de l'analyse des productions d'apprenants en français langue étrangère, de l'évaluation de leurs compétences et de la programmation de l'enseignement: regards sur un domaine de la didactique du français". En: Durand, J., Habert, B., Laks, B. (eds.): *Congrès Mondial de Linguistique Française, Paris, Institut de Linguistique Française*, pp. 621-631. [En línea]: <http://www.linguistiquefrancaise.org/index.php?option=com_article&access=standard&Itemid=129&url=/articles/cmlf/pdf/2008/01/cmlf08318.pdf > [Consulta: 26/03/2010].

VÉRONIQUE, D. (dir.) (2009): *L'acquisition de la grammaire du français, langue étrangère.* Collection Langues & Didactique. Paris: Didier.

VÉRONIS, J. (coord.) (2004): *Le traitement automatique des corpus oraux*. TAL, Traitement automatique de langues, volume 45, nº2/2004. Cachan: Lavoisier/ATALA.

VOGEL, K. (1995) : *L'interlangue, la langue de l'apprenant*. Toulouse : Presses Universitaires du Mirail.

WEBER, C. (2006): "Pourquoi les français ne parlent-ils comme je l'ai appris". En : *Le Français dans le Monde*, nº 345, mai-juin 2006. [En línea]: <http://www.fdlm.org/fle/article/345/weber.php> [Consulta: 07/10/2007].

WHITE, C. (2008): "Beliefs and good language learners." En: Griffiths, C. (dir.): *Lessons from Good Language Learners*. Cambridge : Cambridge University Press, pp. 121-130.

WIDDOWSON, H. G. (2000): "On the Limitations of Linguistics Applied". En: *Applied Linguistics* 21/1, pp. 3-25.

APÉNDICE A:

MUESTRAS DE TRANSCRIPCIÓN DEL CORPUS CORAF

NIVEL A1: A1W03

[
@Title: A1W03
@File: A1W03
@Participants: SUB, Sxxx (woman, C, 2, bank clerck, participant, Paris)
ENT, Ana (woman, B, 3, researcher, participant, Ciudad Real)
@Date: 20/05/2010
@Place: Toledo
@Situation: interview in a classroom, not hidden, researcher observer
@Topic: personal information, hobbies, languages, places
@Level: A1
@Languages_spoken: Spanish (L1), English (A1)
@Time_studying_French: 7 months
@Time_in_French-speaking country:
@Source: CORAF
@Length: 11'16"
@Words: 1255 (Learner: 603)
@Acoustic_quality: A
@Transcriber: A. Valverde
@Revisor: L.Campillos
@Comments: There is some noise from other classrooms and from a building work in progress
]

*ENT: alors bonjour encore ///
*SUB: bonjour ///
*ENT: &eh / comment tu t' appelles ?
*SUB: je m' appelle Sxxx ///
*ENT: et Sxxx / vous → [/] vous → êtes née ici ?
*SUB: non je suis née à → / Paris ///
*ENT: à Paris {%com: ENT also shows surprise} ?
*SUB: <oui / hhh {%act: laugh}> ///
*ENT: [<] <et qu' est ce que vous pouvez me raconter de ça@oral> ?

*SUB: &mm / &eh / mon père et ma mère / &eh / ont → allés → à Paris / pour travailler // et → / je suis née → / là ///
*ENT: et après → / vous êtes venus ici ? <ou / &eh> ...
*SUB: [<] <oui> /// &eh / quand je → [/] j' ai → / deux ans // <&eh> +
*ENT: [<] <hhh {%act: assent}> ///
*SUB: ¬ &eh / eux / &eh / <&mm {%com: SUB asks for the pronunciation of 'eux'}> +
*ENT: [<] <oui eux> ///
*SUB: eux / &eh / volvieron@spa xxx / &mm / &eh / sont venus / &eh / à l' Espagne // et → / je suis → / &eh / ici ///
*ENT: tout le temps à Txxx ?
*SUB: tout@g el@spa temps@g ///
*ENT: d'accord /// et → / qu' est ce que vous faites / comme métier ? comme <travail> ///
*SUB : [<] <comme> travail ? je → [/] je suis → / &eh / à la banque // je travaille à la banque ///
*ENT: d'accord /// et → vous avez étudié → / quelque chose ?
*SUB: hhh {%act: click} / je &s [/] &eh / je étudie / &eh / &mm / tourisme // mais je ne le → / fini pas /// hhh {%act: click} / et → j' ai → / rencontré le → travail / et je → [/] je travaille à la banque <&eh> +
*ENT: [<] <à la banque> ///
*SUB: ¬ pendant / vingt ans ///
*ENT: ah! ben c' est → [/] <c' est déjà>
*SUB: [<] <hhh {%act : assent}> ///
*ENT: ¬ pas mal // <vingt ans> ///
*SUB: [<] <hhh {%act : laugh} / oui> ///
*ENT: et c' est un travail / qui vous plaît ?
*SUB: oui /// je → / &mm [/] j' aime → &eh [/] j' aime → / hhh {%act: doubt} / parler avec la → [/] les personnes // et → / oui / je l' aime /// <hhh {%act : assent}> ///
*ENT: [<] <d'accord> /// et → &mm [/] et qu' est ce que vous faites / pendant &votr [/] votre temps libre ? qu' est ce que vous faites comme / loisirs ?
*SUB: &mm [/] &eh je → [/] j' aime / &eh / me promener / &eh / à Txxx /// &eh / j' adore / &eh / le cinéma / hhh {%act: laugh} // et → [/] et je → [/] je suis → avec ma famille / &eh / avec mes filles ///
*ENT: d'accord // bon on@oral sépare le cinéma // et → / est ce que vous → regardez des films / en français ?

*SUB: en français non ///
*ENT: et vous connaissez [/] ou bien vous connaissez un film en français que → [/] que vous aimez ? même si &l [/] vous l' avez vu en espagnol / c' est pas {%oral: 'ne' absent} grave ///
*SUB: non / je ne xxx vu / hhh {%act: laugh} ///
*ENT: ben c' est normal // le cinéma français c' est pas {%oral: 'ne' absent} très connu → ici // c' est pas {%oral: 'ne' absent} très répandu / <ici donc>
*SUB : [<] <hhh {%act : assent}> ///
*ENT : ¬ c' est normal /// et / est ce que vous aimez lire ?
*SUB: &eh / &mm +
*ENT: la lecture / est ce que vous l' aimez ?
*SUB: ah! oui /// &eh / j@g' aime@g le → [/] &eh / la@g lecture@g / et → je lire@c {%com: je lis} / &eh / trop ///
*ENT: et qu' est ce que vous lisez ?
*SUB: &mm [/] &l [/] des → [/] je → lisez@c {%com: je lis} de [/] &bue [/] je lis / &eh / des → / &mm [/] &mm / ¡ay!@spa / novelas@spa ?
*ENT: des → romans ///
*SUB: ¬ des@g romans@g // hhh {%act: click} / et → [/] et maintenant je suis → [/] &eh / lis / &eh / &mm [/] je suis en train de lire // &per [/] pardon [///] je suis en train de lire / un → premio@spa Nobel / &eh [/] &eh / Sigrid // &eh / c' est → a@eng / &mm / écrivain / &eh [/] &no [/] &norueg [/] <&eh / &n> [/]
*ENT: [<] <norvégien> {%com: ENT speaks slowly and with syllabification for SUB's comprehension} ///
*SUB: ¬ &no [/] &norue [/] <de Noruega@spa / hhh {%act: laugh}>
*ENT: [<] <norvégien> ///
*SUB: [<] <norvégien@g> ///
*ENT: ben c' est difficile hein? / <hhh {%act: laugh}> ///
*SUB: [<] <hhh {%act: laugh}> ///
*ENT: et c' est bien ?
*SUB: oui / <je → [/] je l' aime> ///
*ENT: [<] <et → / de quoi> s' agit il ?
*SUB: &eh / &mm / il &s [/] &eh [/] &eh [/] il est → / sur / &eh / un homme / et → sa famille / et → toute le → [/] la vie de → [/] de sa famille // un peu triste / <et> → ...
*ENT: [<] <hhh {%act: assent}> /// mais au moins si c' est bien / au <moins> ?
*SUB: [<] <oui oui> / <hhh {%act: laugh}> ///

*ENT: [<] <d'accord> /// et → / hhh {%act: click} / vous avez des filles / des fils ?
*SUB: &eh / j'ai des filles // deux ///
*ENT: deux / d'accord /// est ce que vous pouvez me dire quelque chose / sur elles ?
*SUB: &eh / oui // &eh / &mm / ils sont / très belles / <hhh {%act: laugh}> ///
*ENT: [<] <hhh {%act: laugh}> ///
*SUB: elles sont → très étudiantes // &eh / &l → [/] la &m [/] la &mille [/] la / majeure / &eh / la fille → +
*ENT: ¬ aînée ///
*SUB: aînée@g ? &mm / elle est étudiante / &eh / à l' université // elle a → / &eh / &dixs [/] dix-neuf / ans // dix-neuf ans // et → / elle est très intelligente ///
*ENT: et → / elle étudie quoi ?
*SUB: hhh {%act: click} / elle@g étudie@g / des → mathématiques et → [/] et → &ingé [/] ingénieur informatique / les <deux> → +
*ENT: [<] <ah! les> <deux> !
*SUB: [<] <les> deux ...
*ENT: ben c' est difficile ///
*SUB: oui / hhh {%act: laugh} / <mais elle est → > +
*ENT: [<] <même temps → > l' informatique → / le → [/] <l' ingénierie> ...
*SUB: [<] <mais elle est très> contente // et → / elle est → / très intelligente // <xxx> +
*ENT: [<] <d'accord> ///
*SUB: elle +
*ENT: et → / elle le fait → à Mxxx // ou elle le <fait → > ?
*SUB: [<] <à Mxxx> // elle es@spa à l' université → / Autónoma ///
*ENT: ah! d'accord / moi je connais // moi aussi <j' étais → > /
*SUB: [<] <hhh {%act: assent}> ///
*ENT: ¬ à l' Autonóma / <hhh {%act: laugh}> ///
*SUB: [<] <hhh {%act: laugh}> ///
*ENT: et → / &mm [/] et l' autre ?
*SUB: l' autre a → &eh / &dixsi [/] dix-six@n / &eh / années // et elle est → / &eh / faire / &eh / la ESO ///
*ENT: d'accord ///
*SUB: xxx ça@oral ...
*ENT: mais il va bien ? elle <fait → > ...
*SUB: [<] <oui oui> / elle va → / très bien ///

*ENT: xxx &ce [/] c' est très important <ça@oral non?>
*SUB: [<] <oui oui> / <hhh {%act: laugh}> ///
*ENT: [<] <d'accord> /// et → / &mm / est ce que vous connaissez de la musique → / française ?
*SUB: <hhh {%act: doubt}> ///
*ENT: [<] <pas beaucoup> ///
*SUB: pas de → [/] rien /// &ah [/] &eh / non / je ne connais → / plus / &mm [/] &eh [/] &ec [/] &e [/] <&n> +
*ENT: [<] <je ne> [///] pas beaucoup <ou / &eh> ...
*SUB: [<] <hhh {%act: assent} / &mm> / je ne <connais> ...
*ENT: [<] <pas> du tout ///
*SUB: pas@g du@g tout@g ///
*ENT: d'accord /// et → par rapport au français / pourquoi est ce que vous avez choisi / &euh / d' étudier le français ici ?
*SUB: &ah / parce que je → [/] je l' aime &el [/] le français // et → [/] et → mon père / il parle → {%oral: dislocation à gauche} un peu français → / aussi /// et / je → [/] je [/] &mm / je veux → / parler / &eh / quelque {%com: syllabification: quel-que} jour / hhh {%act: laugh} / un peu → meilleur ///
*ENT: d'accord /// et → le français / pour vous c' est / facile // c' est difficile // c' est → / &e [/] pareil → à l' espagnol ? ou qu' est ce que vous pensez du français ?
*SUB: &eh / &mm / pour moi / je → el@spa / hhh {%act: click} [///] l' oral / &eh / pour moi c' est difficile /// &mm / &eh / &l → [/] la → gramatique@n / &mm / c' est difficile / mais je → le → [/] le fais → / bien ///
*ENT: hhh {%act: assent} / d'accord ///
*SUB: mais le [/] l'oral ...
*ENT: xxx mais l' oral / mais aussi la compréhension // <ou &eh>
*SUB: [<] <non / la &comp> +
*ENT: ¬ [<] <seulement> la [/] la façon [/] l' expression ?
*SUB: l' expression@g ///
*ENT: ah! <d'accord> ///
*SUB: [<] <plus> difficile ///
*ENT: hhh {%act: click} / c' est le plus difficile / non c' est <vrai> //
*SUB: [<] <hhh {%act: assent}> ///
*ENT: ¬ c' est → [///] on@oral n' est pas habitués peut-être à parler donc c' est → [/]
*SUB: oui ///

*ENT: ¬ c' est ça /// et → / par rapport aux cours / que vous → / suivez ici est ce que vous êtes contente de la façon de faire les cours ? ou → vous mettrez / des choses différentes // ou → / hhh {%act:doubt} / quelque chose qui manque ?
*SUB: non je suis &tr [/] très contente // &nos [/] nous parlons / &eh [/] &eh / souvent ///
*ENT: hhh {%act:assent} ///
*SUB: ¬ et tout les jours / &eh / nous parlons à &l [/] à la classe // et → / je suis très contente // <parce que si → > /
*ENT: [<] <d'&a> +
*SUB: ¬ tu ne → / parles pas / &eh / tu ne → [/] tu ne sais / &eh / parler ///
*ENT: non {%alt: nan} c' est vrai /// <il faut pratiquer> /
*SUB: [<] <hhh {%act: assent}> ///
*ENT: ¬ <pour →>
*SUB: [<] <il@g faut@g / pratiquer@g> ///
*ENT: ¬ [/] pour pouvoir → améliorer un peu plus // c' est une bonne → / chose // c' était vrai /// et → par rapport au / &mm / hhh {%act: click} = moi je connais pas {%oral: 'ne' absent} Txxx // donc / est ce que / vous pouvez me proposer quelque chose à faire ici / ce soir ?
*SUB: oui / il y a → des → [/] des musées // &eh / il y a → / &eh [/] &l [/] la cathédrale // et → / &mm / vous pouvez → aller maintenant à le musée de → [/] &eh / &mm / Sxxx Cxxx // il y a une exposition / très bon // et → / même aussi le → / &mm / # &eh / Vxxx Mxxx / &eh / il est → / autre → musée ///
*ENT: hhh {%act:assent} ///
*SUB: et → &mm / hhh {%act:click} / &eh / aussi vous pouvez → / promener &a [/] pour les → / pasées@n [/] &eh / les → / hhh {%act:click} / &l [/] les rues //
*ENT: <hhh {%act:assent}> ///
*SUB: ¬ [<] <les rues> de → [/] de Txxx // &eh / c' est très agréable // &mm +
*ENT: ben c' est [/] c' est bien ///
*SUB: sí@spa ///
*ENT: et → moi je veux aussi dîner donc / il y a → quelque chose / bon ou quelque → [/] quelque endroit / quelque restaurant / où je peux → / dîner ce soir // et qui soit bien ?
*SUB: &eh / bien / &eh / il y a → / &eh / le restaurant / qu' il s' appelle / Axxx /// il est → un peu → / cher / mais → tu vas → / à → manger / très <bien> ///

*ENT: [<] <très> bien ///
*SUB: hhh {%act:assent} ///
*ENT: et il se trouve où ?
*SUB: il@g se@g trouve@g / &ah [/] &eh / après → la cathédrale // et → / un peu → / &eh [/] &al@spa [/] &eh / à côté de la → / place del@spa &ayuntamien [/] del@spa → +
*ENT: de la <mairie> ///
*SUB: [<] ¬ <de@g la@g mairie@g> // de la mairie / hhh {%act:laugh} // pardon /// et c' est → très [/] très → &mm # &mm / c' est → / près ici ?
*ENT: c' est près d' ici / <c' est très proche> ///
*SUB: [<] <près d'@g ici / oui> c' est très <proche@g> ///
*ENT: [<] <d' accord> / donc je n' aurai pas de problèmes pour le trouver <non?>
*SUB: [<] <hhh {%act:assent}> /// oui ///
*ENT: d'accord /// et → / &mm [/] et vous aimez cuisiner ?
*SUB: non ///
*ENT: <non {%com: ENT also shows surprise}> ?
*SUB: [<] <c' est> mon mari qui → / cuisine ///
*ENT: ben c' est bien <non?> //
*SUB: [<]<hhh {%act: laugh}> / <oui hhh {%act:laugh}> ///
*ENT: ¬ [<] <comme ça@oral → / on@oral [/] on@oral se décharge> // <c' est lui qui le>
*SUB: [<] <sí@spa> ///
*ENT: ¬ <fait> // <et il a → >
*SUB: [<] <c' est lui@g → / qui@g le@g fait@g> ///
*ENT: ¬ un spécialité / qui fait très bien / ou / &euh / <il fait tout bien> ?
*SUB: [<] <il@g fait@g → / très@g bien@g> / tout /// &eh / il cuisine / &eh / de paella@spa / de → [/] de → omelette / &eh / de → toutes les → / choses ///
*ENT: ben c' est bien ///
*SUB: oui oui / <hhh {%act:laugh}> ///
*ENT: [<] <parce que c' est pas> {%oral: 'ne' absent} / très frequent ça@oral ///
*SUB: non / c' est ...
*ENT: ben c' est une chose bien /// et → pour finir / est ce que vous pouvez me décrire / une personne que / vous aimez bien // que → / bon / que &soi [/] soit intéressante pour vous / ou ?

*SUB: &eh / oui / &eh / ma fille → / &eh / c' est / &eh [/] a → [/] elle est → / blonde /// elle a → &l [/] elle est → / &mm / petite {%alt: petit} // n' est pas grande /// &eh [/] &mm / elle est mince /// et → / elle a → / les yeux bleus [/] bleus {%com: SUB repeats the word for a better pronunciation} /// et → / elle est très sympa@oral /// et → # c' est tout / hhh {%act: laugh} ///
*ENT: ben c' est bien ///
*SUB: hhh {%act: assent} ///
*ENT: alors / c' est suffisant donc on@oral a fini /// xxx je vous remercie beaucoup / parce que → / c' est très important ///

NIVEL B2: B2W02

[
@Title: B2W03
@File: B2W03
@Participants: TAG, Txxx (woman, A, 3, student, participant, San Sebastián)
ENT, Ana (woman, B, 3, researcher, participant, Ciudad Real)
@Date: 18/03/2010
@Place: Facultad de Letras de Ciudad Real (UCLM)
@Situation: interview in a classroom, not hidden, researcher observer
@Topic: personal information, hobbies, languages, places
@Level: B2
@Languages_spoken: Spanish (L1), English (C1), German (A2), Chinese (A1), Arabic (A1).
@Time_studying_French: 8 years (high school and 2 years at university)
@Time_in_French-speaking country: Stages in summer (2 months in Lille and Nantes)
@Source: CORAF
@Length: 15'48"
@Words: 2423 (Learner: 1417)
@Acoustic_quality: A
@Transcriber: A. Valverde
@Revisor: C. Sanz
@Comments:
]

*ENT: alors / bonjour !
*TAG: <bonjour> ///
*ENT: [<] <&eh> / tu peux → / nous dire // comment tu t' appelles → // quel est ton lieu d' origine → // qu' est ce que tu fais → / ici → // bon / ce type de choses / que tu dirais // à une personne → / qui → ne te connaît pas ///
*TAG: &bo [/] &m [/] moi → je suis Txxx /// &eh / j' habite à Cxxx Rxxx // j' étude@c {%com: j'étudie} / de → / la philologie → / française / et anglaise /// &mm / je suis née à → / Sxxx Sxxx // au nord de l' Espagne /// mais → / j' habite dans → / un petit village / de → Cxxx Rxxx // avec ma mère /// et → / bien [///] j' étude → / ici / dans la cité [/] dans la ville //
*ENT: <hhh {%act: assent}> ///

*TAG: ¬ [<] <dans une résidence> universitaire ///
*ENT: et c' est bien / la vie → / à la résidence universitaire ?
*TAG: oui / mais → / quelquefois c' est → [/] c' est un peu → / &mm / &eh / ennuyant // il y a → / &mm [/] &m [///] les → [/] les gens / font des fêtes // et → / <il n' y a pas → > [/]
*ENT: [<] <c' est bruyant> / <peut-être> ///
*TAG: ¬ [<] <il n' y a> pas → / un ambiant d' étude // vraiment ///
*ENT: ah! d'accord /// et / il y a / des gens / de ton village // aussi dans la résidence // ou pas ?
*TAG: oui // &eh / il y en a → / trois → / ou quatre /// <hhh {%act: assent}> ///
*ENT: [<] <hhh {%act: assent}> /// alors c' est bien non? &auv [/] au moins → ...
*TAG: <hhh {%act: assent}> ///
*ENT: [<] <et tu ne penses> pas → / peut-être → / déménager à → / un appartement → / avec tes collègues ou → ?
*TAG: oui peut-être la → / prochaine année // &eh / parce que / je crois que → / j' aurais plus / d' &inde [/] d' indépendance → et → / je pourrais / faire les choses / quand → / je → voudrais ///
*ENT: hhh {%act: assent} ///
*TAG: ici / dans la résidence / je dois → / &mm [/] &mm [/] hhh {%act: doubt} [/] m' &a → [///] je dois [/] <j' ai mes horaires → > /
*ENT: [<] <respecter> / plutôt <les horaires> ///
*TAG: ¬ [<] <je> = oui /// mais → / j' ai aussi des avantages // je ne dois / préparer la nourriture → / <et tout ça> ///
*ENT: [<] <c' est vrai> ///
*TAG: et → / c' est de [/] du temps / que → je profite / pour moi /// <hhh {%act: assent}> ///
*ENT: [<] <et c' est vrai> /// et alors → / &eh / bon // &eh / &mm / en parlant de tes études // pourquoi est ce que tu as choisi le → [/] la philologie → / anglaise / ou française ?
*TAG: bon / dès → que j' étude@c {%com: j'ai étudié/ j'étudiais} / au lycée // j' ai aimé → / les langues // surtout le français /// mais je ne peux pas exprimer pourquoi → // j' aime sa → sonorité / c' est une langue → / très belle {%alt: belt} [/] belle // et aussi → / quand → j' ai connu la littérature // je l' ai aimée beaucoup // et aussi je → [/] j' aime étudier de l' anglais // parce que c' est très important / mais → / je préfère la → [/] la philologie française ///
*ENT: d'accord /// <c' est bien> //

*TAG: [<] <hhh {%act: assent}> ///
*ENT: ¬ c' est [///] moi j' aime bien //
*TAG: <hhh {%act: laugh}> ///
*ENT: ¬ [<] <ce type de> <choses> ///
*TAG: [<] <oui> ///
*ENT: et / alors tu aimes lire non? / je <suppose> ///
*TAG: [<] <hhh {%act: assent}> ///
*ENT: et qu' est ce que tu aimes lire normalement ?
*TAG: &mm / j' aime@g lire@g / &eh / un peu de → [/] de tous → / les &gen [/] genres // mais → / j' aime surtout la → / poésie ///
*ENT: ah! c' est <bien> //
*TAG: [<] <oui> ///
*ENT: c' est pas {%oral: 'ne' absent} normal ça ///
*TAG: non // hhh {%act: laugh} ///
*ENT: hhh {%act: laugh} ///
*TAG: je ne sais pas // c' est un → [/] c' est plus littéraire // et → / plus → sentimentale / quelquefois → // hhh {%act: assent} /// <j' aime bien> ///
*ENT: [<] <et → / est ce que tu as> un auteur → / préféré // par exemple en parlant de poésie ?
*TAG: oui // &eh / Rimbaud // <je l' aime beaucoup> ///
*ENT: [<] <hhh {%act: assent}> ///
*TAG: et → aussi les → / symbolistes // à partir de / Baudelaire //
*ENT: <aha!> ///
*TAG: [<] <je les aime> bien // xxx ///
*ENT: et quant aux espagnols → // qu' est ce <que tu pourrais nous dire> ?
*TAG: [<] <&mm> / &l [/] les auteurs de → / la génération du → / vingt sept //
*ENT: <hhh {%act: assent}> ///
*TAG: ¬ [<] <avec> Lorca → // Cernuda ///
*ENT: hhh {%act: assent} /// c' est bien / c' est bien /// &eh / et tu connais → / aucun &e [/] écrivain → / &eh / français // contemporain ?
*TAG: contemporain@g ? &mm / je@g connais@g / &m [///] on@oral a → / li@c {%oral: on a lu} / par exemple → / &mm / qui a écrit → / La Dame Rose [/] Oscar et la Dame Rose // Schmitt [/] <Schmitt → > //
*ENT: [<] <ah! hhh {%act: assent}> /// <Éric Emmanuel / Schmitt> ///

*TAG: ¬ [<] <&eh / &mm> / Guillaume Musso // ce sont les plus → +
*ENT: les plus connus / peut-être ///
*TAG: hhh {%act: doubt} // oui /// les plus → / actuels //
*ENT: <hhh {%act: assent}> ///
*TAG: ¬ [<] <peut-être> /// <hhh {%act: laugh}> ///
*ENT: [<] <c'est> bien /// et → / bon / &eh / tu as déjà visité la France ?
*TAG: oui /// &eh / &mm / grâce aux → bourses → / de l' été //
*ENT: <hhh {%act: assent}> ///
*TAG: ¬ [<] <de la → > / Junta de Castilla La Mancha {%com: regional governement} / et → du ministère // j' ai été → / à Lille // pendant / un mois //
*ENT: <hhh {%act: assent}> ///
*TAG: ¬ [<] <avec> une famille [/] une famille // et → / &a [/] une [/] un autre mois → / à Nantes // avec une autre famille ///
*ENT: et ça a été bien l' expérience ? qu' est ce que tu peux nous raconter à propos de l' expérience → // <d' être là → > //
*TAG: [<] <oui> ///
*ENT: ¬ dans une famille / française ?
*TAG: hhh {%act: assent} /// je → [/] j' ai profité beaucoup /// j' ai appris → [/] j' ai amélioré → / mon français // et → / j' ai vi@c {%com: j' ai vécu} [/] j' ai vécu / ce que → [/] ce [/] ce qu' on@oral → / vit en → [/] en France // &eh / &mm / j' ai été avec la famille // dont j' ai appris / leurs / cotumes@n → et tout ça@oral /// le matin j' allais → / à l' école // <avec d' autres> /
*ENT: [<] <hhh {%act: assent}> ///
*TAG: ¬ espagnols / c' est la → [/] la part que je ne → [/] je n' ai pas aimée <de → > +
*ENT: [<] <que tu> <regrettes> ///
*TAG: [<] <hhh {%act: laugh}> / oui /// hhh {%act: assent} ///
*ENT: et → / qu' est ce que tu penses de la France ?
*TAG: &mm / hhh {%act: laugh} ///
*ENT: c' est compliqué // mais bon ...
*TAG: oui // c' est un peu général // <je ne sais pas → / hhh {%act: laugh}> ///
*ENT: [<] <oui> ///
*TAG: &m [/] &ah [/] &mm [/] au premier → regard / c' est → [///] bien que c' est un pays très proche // il y a des → / &mm [/] des

différences → / culturelles / par exemple je → [/] je considère que la culture française // et l' art → / &eh / est [/] a plus / de vitalité qu' en Espagne // les → [/] les → jeunes / &eh [///] il y a des expositions → // et des &m [/] &mus [/] musées → / gratuits / pour les gens //
*ENT: <hhh {%act: assent}> ///
*TAG: ¬ [<] <et → > / <&mm> +
*ENT: [<] <il y a un mouvement> culturel / plus → / approfondi qu' <ici peut-être> ///
*TAG: [<] <oui> /// hhh {%act: assent} /// surtout → à Paris // <je crois> ///
*ENT: [<] <hhh {%act: assent}> /// oui / c' est bien /// et → / tu aimerais → / peut-être vivre en France // ou travailler / pendant un temps // o@spa tu penses → / voyager → / hhh {%act: inhalation} = je ne sais pas / bon / peut-être pas voyager / mais → / &eh / vivre là comme Erasmus // <hhh {%act: laugh}> ?
*TAG: [<] <oui> ! hhh {%act: laugh} /// l' année prochaine / je → [/] je serai là // <en Erasmus> ///
*ENT: [<] <ah! oui ? tu seras où> ?
*TAG: à Paris ///
*ENT: ah! pas mal !
*TAG: oui ! {%com: TAG laughs while she is speaking} c' est génial ! <hhh {%act: assent}> ///
*ENT: [<] <oui> // et → / qu' est ce que tu penses faire / quand tu sois là ?
*TAG: &mm +
*ENT: tu vas être avec les espagnols <tout le temps> //
*TAG: [<] <non non> <non non> //
*ENT: ¬ [<] <ou non> ? <hhh {%act: laugh}> ///
*TAG: ¬ [<] <je ne veux pas ça@oral> /// &m [/] maintenant je suis en train de → / chercher → / une résidence // parce que → / il y a beaucoup → / des gens qui → [/] qui la sollicitent //
*ENT: hhh {%act: assent} ///
*TAG: ¬ mais → / surtout étudier // c' est vrai // mais aussi → / avoir une vie culturelle // &eh / et tout ça@oral /// <hhh {%act: assent}> ///
*ENT: [<] <et connaître> / la culture un <peu → / à fond / non? > ///
*TAG: [<] <hhh {%act: assent}> ///
*ENT: c' est bien /// &eh / &mm / on@oral va → / changer un peu de <sujet / alors → > //

*TAG: [<] <o kay@eng> ///
*ENT: ¬ qu' est ce que tu penses faire → // à l' avenir → // &eh / quand tu finis tes études ?
*TAG: j' ai → / toujours voulu être → / professeure //
*ENT: hhh {%act: assent} ///
*TAG: ¬ au lycée // mais → en Espagne → // j' ai vu / qu' il n' y a pas de professeurs de → / littérature / parce que j' aime plus / la littérature <que → / la langue> //
*ENT: [<] <que la langue> ///
*TAG: ¬ &s [/] donc → / au lycée je ne pourrais → / pas être → / professeure <je crois> +
*ENT: [<] <ben peut-être> à la fac@oral ///
*TAG: oui // <j' aimerais → > /
*ENT: [<] <hhh {%act: laugh}> ///
*TAG: ¬ ça@oral // <hhh {%act: laugh}> ///
*ENT: [<] <hhh {%act: laugh} / aha! d'accord> ///
*TAG: étudier → / plus /// peut-être un → doctorat ///
*ENT: hhh {%act: assent} ///
*TAG: et → / donc devenir professeure de → [/] de faculté [/] d' université ///
*ENT: hhh {%act: assent} /// c' est bien /// et → / on@oral change encore // ça@oral c' est → / tout le temps comme ça@oral /// hhh {%act: laugh} /// et → / quel est le dernier film / que tu as → / regardé à la télé / ou → &eh [///] je ne sais pas / peut-être au cinéma // si tu vas / souvent au cinéma ?
*TAG: hhh {%act: doubt} ///
*ENT: ou le type de film / que tu aimes / voir → // <normalement> ?
*TAG: [<] <&bue [/] la → / dernière [/] le dernier film que → / j' ai vu / xxx [/] le cinéma // c' est → / Une Éducation // c' est une → [/] c' est un film / anglaise [/] anglais /// <&eh> +
*ENT: [<] <je voudrais le voir> // un jour // hhh {%act: assent} ///
*TAG: hhh {%act: laugh} /// &eh / qui parle / sur → une fille // à → Londres // qui est → [/] qui est une → / très bonne / étudiante // et → tout d' un coup / elle [/] elle connaît → / un homme / plus → [/] plus âgé → / qu' elle // qui → lui fait / &decouvr [/] découvrir → / la vie artistique → // les → concerts de jazz // donc elle → [/] elle se trouve → / divisée → // entre ses études // et → / cette vie // d' amour et → / de musique / et tout ça@oral ///
*ENT: hhh {%act: assent} /// et c' est bien ?

*TAG: oui oui ///
*ENT: tu me le conseilles ?
*TAG: oui // <hhh {%act: laugh}> ///
*ENT: [<] <c'est bien> /// et tu connais des → / films français → // ou → / tu les regardes // souvent des films français ?
*TAG: oui / j' aime / voir des → [/] des films en → / version originale pour → [/] pour écouter /// et → / la plus connue peut-être / Amélie {%com: normally known in French as Amélie Poulain} //
*ENT: hhh {%act: assent} ///
*TAG: ¬ et → aussi → / Jeux d' Enfants // <et → > +
*ENT: <aha! ça@oral bien> // avec → / Marion → / <Cotillard> /
*TAG: [<] <oui> /// <hhh {%act: laugh}> ///
*ENT: ¬ [<] <et Guillaume / Canet> /// <hhh {%act: laugh}> ///
*TAG: [<] <hhh {%act: laugh}> ///
*ENT: moi j' aime / Guillaume Canet //
*TAG: hhh {%act: laugh} /// <moi aussi> ///
*ENT: ¬ [<] <donc c' est pour ça> / que le je sais / très bien /// hhh {%act: laugh} /// et par rapport à de → [/] à de la musique [///] bon à &le [/] à la musique française → // est ce que tu écoutes / de la musique / française ?
*TAG: oui /// les → premiers / musiciens / que j' ai écoutés en français / ce sont → / Edith Piaf → [///] &mm / grâce à ma mère // parce que ma mère aimait le français aussi // et elle m' a fait → / écouter → / &mm / Brassens → / Léo Ferré → // et tous les → / <classiques> /
*ENT: [<] <tous les grands> / <on@oral va dire> //
*TAG: ¬ [<] <on [/] on peut> <dire /// oui> ///
*ENT: ¬ [<] <de la musique> ///
*TAG: et → / hhh {%act: doubt} / des groupes → / plus modernes peut-être → / &mm [/] &mm # Noir Désir → # Superbus → ...
*ENT: hhh {%act: assent} ///
*TAG: mais → / j' aime plutôt les → [/] <les grands> //
*ENT: [<] <les anciens> // <on@oral peut dire> ///
*TAG: [<] <oui> // <oui> ///
*ENT: [<] <hhh {%act: laugh}> ///
*TAG: <hhh {%act: assent and laugh}> ///
*ENT: [<] <et → > / tu as une chanson préférée // que tu pourrais nous conseiller // ou → ?
*TAG: oui de → / Boris Vian → //
*ENT: <hhh {%act: assent}> ///

*TAG: ¬ <Le Déserteur> ///
*ENT: et pourquoi tu aimes bien / cette → / chanson ?
*TAG: pour → [/] pour le sujet // de la guerre // d' une personne que → [/] qui ne veut → / pas faire la guerre // hhh {%act: assent} // et qui / écrit au Président → / lui disant // si tu veux → / lutter → / pour [/] pour ton pays → // va → / &t [/] &tui [/] toi même ///
*ENT: <hhh {%act: assent}> ///
*TAG: [<] <hhh {%act: assent}> ///
*ENT: d'accord /// c' est bien / c'est bien /// et → / qu' est ce que tu vas faire / pendant les vacances // qui → / sont déjà → / <très proches> ?
*TAG: [<] <oui> /// bon / en → juillet //
*ENT: hhh {%act: assent} ///
*TAG: ¬ &e [/] j' irai au → [/] au Caire // en Égipte ///
*ENT: oh! <pas mal> !
*TAG: [<] <oui> /// &eh / il y a deux jours que → / j' ai reçu la nouvelle // que → / &j [/] &ji [/] j' y serai // <en juillet> ///
*ENT: [<] <hhh {%act: assent}> ///
*TAG: et → c' est pour faire un → [/] un cours d' arabe // <avec → > +
*ENT: [<] <ben tu aimes aussi → > / &eh / <l' arabe> ?
*TAG: [<] <oui> / le → [/] toutes les langues //
*ENT: <toutes le langues // c' est bien> ///
*TAG: ¬ [<] <j' aime étudier / les langues> /// oui ///
*ENT: c' est → le [/] le [/] la philologue / parfaite <alors> ///
*TAG: [<] <hhh {%act: laugh}> ///
*ENT: hhh {%act: laugh} ///
*TAG: et → / en → août {alt: u} // je ne sais pas encore // peut-être je → travaillerai → // &mm / chez moi // dans mon village // en → / donnant des cours → / ou → je ne sais pas /// et → / donc / en &se [/] en → / septembre // je dois → / &pre [/] me préparer pour aller → / en Erasmus /// <hhh {%act: assent}> ///
*ENT: [<] <c'est bien> / c' est bien /// de très bons projets // <je pense → > ///
*TAG: [<] <hhh {%act: assent}> ///
*ENT: et / &eh / &mm / bon on@oral parlé de la musique // on@oral a parlé de → [/] de [/] de ce qui est lire aussi → /// &eh / maintenant → / est ce que tu peux → / nous dire → // est ce que tu aimes faire du sport ?

*TAG: j' aime@g → / faire@g du@g sport@g mais → / maintenant / à ce moment là je n' ai pas → / le temps de → [/] de faire // <vraiment> ///
*ENT: [<] <aha! > / et tu as pratiqué / quelque sport / <en → > ...
*TAG: [<] <oui / quand> j' étais au lycée // j' étais dans → / une équipe de → [/] un équipe de → / basket /// parce que j' étais → / grande // <donc → > /
*ENT: [<] <hhh {%act: laugh}> ///
*TAG: ¬ mais je n' étais pas / très bonne ///
*ENT: non ?
*TAG: je dois → / le dire /// <hhh {%act: laugh}> ///
*ENT: [<] <hhh {%act: laugh}> ///
*TAG: non mais → / j' ai toujours aimé → / nager ///
*ENT: aha! // <très bien> ///
*TAG: [<] <nager /// hhh {%act: assent}> ///
*ENT: et → / pour ce qui est / de → / &mm [/] de ton repas préféré → // aimes tu cuisiner → // ou → [/] ou / tu as → / un repas → / spécial → / que tu → fais / <à tes amis → > /
*TAG: [<] <hhh {%act: assent}> ///
*ENT: ¬ quand ils sont / chez toi → // ou → / hhh {%act: laugh} ?
*TAG: &mm / je ne / cuisine pas → trop // parce que → {%com: ENT laughs} / comme j' avais dit → avant // je suis dans une résidence // mais → / quand je suis chez moi // j' essaie de → [/] d' aider ma mère // ou de préparer → / des choses // je → / &mm [/] je préfère → / cuisiner quelque chose de nouveau ///
*ENT: <hhh {%act: assent}> ///
*TAG: [<] <par exemple → > / la dernière chose que → / j' ai cuisinée / c' est un → [/] un repas → / allemand //
*ENT: <ah!> ///
*TAG: ¬ [<] <que j' ai> lu dans → [/] dans [/] dans un livre // de cuisine ///
*ENT: c' est bien ///
*TAG: donc → / des → choses / nouvelles /// hhh {%act: assent} ///
*ENT: pas mal // donc → / si tu vas en France // tu vas apprendre // peut-être / à faire → / des choses françaises / que tu pourras peut-être / &eh / faire goûter // <hhh {%act: laugh}> /
*TAG: [<] <hhh {%act: assent and laugh}> ///

*ENT: ¬ au reste du monde /// et → / qu' est ce que tu penses → / des cours ici à la fac@oral ? tu aimes bien la façon de donner les cours // ou tu penses qu' il faudrait → / changer quelque chose → // ou → / qu' est ce que tu ferais // comme prof@oral ? comme tu as → l' esprit / d' être professeure du [/] &m [///] bon de langues // ou de littérature à la fac@oral // qu' est ce que tu ferais ?
*TAG: bon / ici j' ai trouvé des professeurs très → / différents // qui ont de → / différentes méthodologies // il y a en a des → [/] il y a des professeurs qui → [/] qui quelquefois / ne viennent pas / ici // et d'autres qui sont très → engagés avec nous // et nous fait → [/] font travailler beaucoup // et je préfère ça@oral /// et → / &mm / par exemple / en littérature / il y a des / professeurs // qui nous donnent / tous les données // et on@oral → [/] on@oral → / écrit ///
*ENT: hhh {%act: assent} ///
*TAG: et → d'autres / qui nous font → / préparer / le sujet // par nous mêmes // et → / faire des expositions // donc → [/] donc / hhh {%act: doubt} / le reste de copains / peuvent nous écouter /// et je préfère &c [/] cette / dernière / méthode // parce que → / c' est plus → active // <et c' est → > /
*ENT: [<] <oui> ///
*TAG: ¬ l' élève [/] le → / étudiant // qui recherche // et qui prépare → / le → [/] le sujet // le livre et tout ça@oral ///
*ENT: c' est une façon de se mêler / <peut-être → / avec> /
*TAG: [<] <hhh {%act: assent}> ///
*ENT: ¬ le sujet plus → [/] d' une façon plus approfondie /// c' est bien /// hhh {%act: click} /// et → / qu' est ce que tu penses / de nouvelles technologies ? tu les utilises / <souvent → > /
*TAG: [<] <oui> ///
*ENT: ¬ <ou → > ?
*TAG: [<] <oui> / le téléphone → / l' Internet → /// maintenant je crois qu' un étudiant ne peut [/] ne peut pas → / être sans Internet ///
*ENT: hhh {%act: assent} ///
*TAG: parce que → / tous les → / travaux → // o@spa → / &mm [/] o@spa les → / hhh {%act: doubt} / hhh {%act: click} = bon / <dans la page → > /
*ENT: [<] <o@spa dossiers → > ///
*TAG: ¬ de → l' université → // <on@oral a des → > [/]
*ENT: [<] <aha!> ///

*TAG: ¬ des dossiers que les professeurs / nous envoient → / et tout ça@oral // &eh / mais → / je n' aime pas / trop la → technologie ///
*ENT: ah! non ? <hhh {%act: laugh}> ///
*TAG: [<] <je ne sais pas / pourquoi> /// non // non ///
*ENT: tu n' es pas très → / douée peut-être ?
*TAG: &mm +
*ENT: o@spa tu n' as pas encore beaucoup → / travaillé avec → / la [/] les nouvelles technologies ?
*TAG: je les utilise // je me débrouille bien mais → // je pourrais → / vivre sans [/] sans &technolo [/] sans technologie je crois /// <hhh {%act: assent}> ///
*ENT: [<] <hhh {%act: assent}> /// et → / tu &re [///] peut-être tu → / regardes sur Internet → / des choses en français ? c' est à dire → // les journaux → // ou → magazines // ou → des blogs // o@spa ...
*TAG: mais [/] plutôt des → vidéos //
*ENT: <des vidéos> ///
*TAG: [<] <ou des chansons> /// oui ///
*ENT: c' est bien /// non mais c' est déjà beaucoup / hein? ///
*TAG: <hhh {%act: assent}> ///
*ENT: [<] <c' est → > [/] c' est pas {%oral: 'ne' absent} /// et → / je crois que pour finir → // tu peux nous raconter quelque chose → / &mm / de toi → // que tu considères → // bizarre / amusante / ou → / caractéristique ?
*TAG: hhh {%act: doubt} ///
*ENT: comment tu pourrais te décrire → / comme ça@oral → = je ne sais pas ?
*TAG: c' est <difficile> ///
*ENT: [<] <c' est> difficile // <mais bon> ...
*TAG: [<] <oui> /// &m [/] tout d'abord / mon → prénom // Txxx parce <que → > /
*ENT: [<] <c' est pas → > {%oral: 'ne' absent} / <très normal> ///
*TAG: ¬ [<] <personne → > / n' a → [/] n' a pas ce → [/] <ce prénom> ///
*ENT: [<] <et quel est l' origine> / de ton nom [/] <de ton prénom> ?
*TAG: [<] <&m [/] mes parents> [/] mes parents l' ont → / inventé ///
*ENT: ah! oui ?
*TAG: oui // hhh {%act: laugh} /// ce [/] <c' était comme un → > [/]
*ENT: [<] <ah! ben c' est génial> !

*TAG: ¬ comme un jeu // chacun disait → / une lettre // xxx non? /// <ils sont [/] ils ont crée → > /
*ENT: [<] <&mm / j' aime bien> !
*TAG: ¬ un mot = oui /// mais → / moi / je ne sais pas /// j' adore / surtout les activités qui → [/] qui ont → / une &relati [/] un rapport avec l' art // j' ai étudié de la musique // <et je → > [/]
*ENT: [<] <ah! sí@spa> ?
*TAG: ¬ je joue de → / la flûte // <&tra> [/]
*ENT: [<] <ah! pas mal> ///
*TAG: ¬ &t [/] la traversien@n // <je ne sais pas> //
*ENT: [<] <moi j' ai joué → > / <de la guitare> //
*TAG: ¬ [<] <flûte> ///
*ENT: ¬ <donc> ///
*TAG: [<] <ah! non> ? très bien ///
*ENT: je connais aussi → / le monde {%com: ENT laughs while she is speaking} / du Conservatoire // <hhh {%act: laugh}> ///
*TAG: [<] <non / je n'ai pas étudié / au Conservatoire> /// hhh {%act: laugh} /// mais → je ne sais pas ///
*ENT: mais tu as → / encore le temps de le faire ///
*TAG: oui // peut-être // mais → avec les → [/] les études // c' est trop compliqué ///
*ENT: oui // je sais ///
*TAG: pour le temps // hhh {%act: assent} ///
*ENT: alors c' est bien /// donc je pense / que ça ira /// oui /// <ça va> [/]
*TAG: [<] <okey@eng> ///
*ENT: ¬ ça va être très bien / donc // et je te remercie beaucoup // <de ta collaboration> //
*TAG: [<] <merci à vous> /// <hhh {%act: laugh}> ///
*ENT: ¬ [<] <d'accord> ?
*TAG: okey@eng ///

NIVEL C1: C1M01

[
@Title: C1M01
@File: C1M01
@Participants: ALB, Axxx (man, C, 3, English teacher, participant, Burgos)
ENT, Ana (woman, B, 3, researcher, participant, Ciudad Real)
@Date: 20/05/2010
@Place: EOI Toledo
@Situation: interview in a classroom, not hidden, researcher observer
@Topic: personal information, hobbies, languages, places
@Level: C1
@Languages_spoken: Spanish (L1) and English (C1)
@Time_studying_French: 4 years
@Time_in_French-speaking country:
@Source: CORAF
@Length: 17'00
@Words: 2416 (Learner: 1464)
@Acoustic_quality: A
@Transcriber: A. Valverde
@Revisor: L. Campillos
@Comments: ALB usually pronounces loudly all final 't' of words. There are some noises at the beginning because the digital recorder falls and ENT tries to fix it, and a continuous noise, too, because ALB beats the table with his fingers while he is speaking. Finally, there are some voices from other classrooms.
]

*ENT: alors bonsoir ///
*ALB: bonsoir ///
*ENT: et → / comment tu t' appelles ?
*ALB: moi je m' appelle Axxx ///
*ENT: Axxx /// et → / vous êtes → / né ici / à Txxx ?
*ALB: non non non / je suis@g → né@g à &Bxx [/] Bxxx ///
*ENT: aha! ///
*ALB: et → / oui oui ///
*ENT: et → / qu' est ce que → / tu peux me raconter à propos de Bxxx ///

*ALB: pues@spa c' est → [/] c' est → [/] c' est une ville / magnifique // et → / je [/] je suis → / né là-bas et → / pour moi c' est la → [/] la meilleure → / <ville> ///
*ENT: [<] <hhh {%act: laugh}> ///
*ALB: ¬ du monde <entier> ///
*ENT: [<] <du monde> ! hhh {%act: laugh} /// et qu' est ce qu' il y a → / à faire ou à voir → à [/] à → / Bxxx ?
*ALB: hhh {%act: click} / &eh / à faire ? <je [/] je> [/]
*ENT: [<] <oui> ///
*ALB: ¬ je suis né → / là-bas et j' ai vécu → / à Bxxx / &eh / pendant → {%com: ALB pronounces final t} treize /
*ENT: hhh {%act: assent} ///
*ALB: ¬ ans // et → / donc je → [/] je [/] je suis allé à Vxxx ///
*ENT: hhh {%act: assent} ///
*ALB: hhh {%act: click} / et j' ai vécu → à Vxxx // je travaillais à Vxxx /// et maintenant je → / &eh [/] j' habite à / Sxxx ///
*ENT: hhh {%act: assent} ///
*ALB: mais → / les derniers → / quatre ans / je [/] je [/] je [/] j' ai vécu à / Sxxx ///
*ENT: aha! et c' était bien / Sxxx ?
*ALB: très différente ///
*ENT: très différent /// <hhh {%act: laugh}> ///
*ALB: [<] <très> différent en comparaison avec ma → [/] hhh {%act: click} [/] ma ville Bxxx / et Vxxx // mais → / xxx importe /
*ENT: <hhh {%act: assent}> ///
*ALB: ¬ [<] <parce que je suis> marié → // et → je me suis marié avec → / une femme à Sxxx // <et donc → >
*ENT: [<] <hhh {%act: assent}> /// <donc il faut &a> +
*ALB: ¬ [<] <je suis venu / &eh> /
*ENT: et oui ! <c' est vrai> ///
*ALB: ¬ [<] <ici> ///
*ENT: il faut → [/] <hhh {%act: laugh}> [/]
*ALB: [<] <hhh {%act: laugh}> ///
*ENT: ¬ il faut y aller → / hhh {%act: laugh} / <là où → il → > [/]
*ALB: [<] <et → [/]> <et &y> [/]
*ENT: [<] <l' &am [/] l' amour> <se trouve> ///
*ALB: ¬ [<] <et je travaille → > à <Sxxx> //
*ENT: [<] <hhh {%act: assent}> ///
*ALB: ¬ [<] <maintenant> ///
*ENT: [<] <d'accord> /// et qu' est ce que → / tu fais comme métier ?

*ALB: hhh {%act: inhalation} / je [/] je suis {%com: syllabification: s-uis} / professeur d' anglais ///
*ENT: aha! ///
*ALB: mais → / hhh {%act: click} / c' est un peu difficile // parce que → / &eh / les élèves en général &eh / hhh {%act: click} / ils / &eh / n' ont pas / un → grande → {%alt: granfe} / motivation //
*ENT: <ah! oui> ///
*ALB: ¬ [<] <por@spa> [/] pour les études <mais → > [/]
*ENT: [<] <hhh {%act: assent}> ///
*ALB: ¬ mais je fais → / &mon [/] mon mieux [/] ma mieux / <por@spa [/] por@spa mes → > +
*ENT: [<] <de mon mieux> /// <hhh {%act: assent}> ///
*ALB: [<] <mon@g mieux@g por@spa> [/] por@spa enseigner l' anglais ///
*ENT: l' anglais / hhh {%act: laugh} /// mais → / ils aiment l' anglais ou → / ils pensent que c' est pas {%oral: 'ne' absent} nécessaire → / ou → ?
*ALB: en général <je crois que → > +
*ENT: [<] <parce que moi> / de [/] de [/] depuis → / ma → [/] ma → profession / xxx [/] depuis le français → // ils disent tout le temps que le français ne sert à rien ///
*ALB: <hhh {%act: assent}> ///
*ENT: [<] <et que c' est> l' anglais // donc → [///] c' est pour ça@oral que je te demande / hhh {%act: laugh} ///
*ALB: &s [/] &s [/] il &peu [/] je crois qu' il y a un grand différence → / là &d [/] &eh / des groupes d' étudiants et → / hhh {%act: click} [///] un groupe d' étudiants qui [/] qui [/] qui est / très très motivé // mais → / par contre il y a / <&eh> /

*ENT: [<] <hhh {%act: assent}> ///
*ALB: ¬ un autre groupe / &eh d' étudiants qui → [/] hhh {%act: inhalation} [/] qui → [/] qui [/] qui ne [/] ne → [/] ne font pas → / un grand effort / <pour → > [/]
*ENT: [<] <hhh {%act: assent}> ///
*ALB: ¬ pour étudier → / mais → ...
*ENT: mais bon ///
*ALB: hhh {%act: click} / mais c' est la <vie> !
*ENT: [<] <mais> <c' est la vie // c' est vrai> //
*ALB: [<] <et → > ...

*ENT: ¬ et il faut le faire /// hhh {%act: laugh} /// et → / &mm [/] et pourquoi est ce que → / tu as choisi le français // pour l' étudier ici à l' école des langues ?
*ALB: parce que → / hhh {%act: blow} / il y a beaucoup {%com: syllabification: beau →-coup} de temps que j' ai étudié → / français à Vxxx /// dès de troisième année ///
*ENT: <hhh {%act: assent}> ///
*ALB: [<] <dans l' école> des → [/] des langues à Vxxx et → [/] et → [/] et cette année → / je [/] je [/] j' ai le temps de [/] de l'étudier / parce que → / j' ai demandé le [/] le → [/] le congé de [/] pour convenance → / <personal@spa>
*ENT: [<] <aha!> ///
*ALB: ¬ <pour → > [/]
*ENT: [<] <hhh {%act: assent}> ///
*ALB: ¬ pour m' occuper de → [/] de mes enfants /// j' ai trois enfants ///
*ENT: ah! oui /// <hhh {%act: laugh}> ///
*ALB: [<] <et un bébé> /// hhh {%act: laugh} {%com: ENT laughs} / donc → / cette année je suis à la maison et → / hhh {%act: click} [/] et → / hhh {%act: inhalation} [/] et je [/] je [/] j' ai la chance de [/] de → [/] de venir ici à Txxx / <et → > [///]
*ENT: [<] hhh {%act: assent} ///
*ALB: ¬ deux jours par semaine / et / pour [/] pour moi c' est → [/] hhh {%act: inhalation} [/] c' est comme une <libération> ///
*ENT: [<] <hhh {%act: laugh}> /// c' est vrai /// alors le français c' est la libération /// <hhh {%act: laugh}> ///
*ALB: [<] <oui> {%com: ENT laughs} ///
*ENT: d'accord /// et → / tu crois que le français c' est → / &eh [///] qu' est ce que c' est pour toi // c' est plus difficile // c' est moins difficile // c' est → / différent // ce → [/] c' est une langue → ... quel &ty +
*ALB: c' est différent // &eh / ce n' est pas difficile si [/] si on@oral le → [/] si on@oral l' étude@c {%com: étudie} / et l' on@oral → [/] hhh {%act: doubt} [///] je crois que c' est [/] que c' est como@spa tout → / dans la vie // si → / on@oral [/] hhh {%act: click} [/] et si on@oral fait un peu d' effort // je crois que c' est → [/] c' est très → / hhh {%act: click} / facile / &eh / de [/] de [/] <de / acquérir> ///
*ENT: [<] <hhh {%act: assent}> /// <d'accord> ///
*ALB: [<] <mais → > / pour moi c' est → [/] c' est une langue qui est / très belle ///

*ENT: hhh {%act: assent} ///
*ALB: en comparaison avec → &eh / l' anglais // <parce que → > /
*ENT: [<] <l' anglais> ? ///
*ALB: ¬ je crois que le &fran [/] &eh [/] le [/] le &f [/] le français c' est / hhh {%act: blow} / plus courtisan <et → >
*ENT: [<] <aha!> ///
*ALB: ¬ plus → / lyrique // <et → > [///]
*ENT: [<] <hhh {%act: assent}> ///
*ALB: ¬ d' après moi /// <mais → > +
*ENT: [<] <hhh {%act: assent}> /// moi c' est peut-être la sonorité aussi / c' est <différente> ///
*ALB: [<] <hhh {%act: assent}> ///
*ENT: donc ça@oral fait → / peut-être une idée → / de → [///] bon la <langue → ... hhh {%act: laugh}> ///
*ALB: [<] <hhh {%act: laugh}> ///
*ENT: comme ça@oral très sonore très belle // hhh {%act: laugh} / <la musicalité → > /
*ALB: [<] <&musi [/] hhh {%act: assent}> ///
*ENT: ¬ donc → ... <oui oui &s> +
*ALB: <je crois que> c' est [/] c' est [/] c' est nécessaire / connaître // bien connaître un langue / pour → [/] pour savoir &eh / à juger de la → / idiosyncrasie ? <de → > [/]
*ENT: [<] <hhh {%act: assent}> ///
*ALB: ¬ de la langue // et de la société française // &eh / je crois ///
*ENT: hhh {%act: assent} /// c' est bien // c' est une bonne idée /// et → / &eh / pour toi → / &eh / est ce que tu as des difficultés // pour apprendre quelque chose en → [/] en français // c' est à dire // &e [/] il y a → / quelque chose qui te coûte un peu plus que le reste ? ou → / c' est tout bien → / et il n' y a pas de difficultés → / ou → ?
*ALB: oh! bien sûr ! le [/] la → [/] la conjugaison verbale je crois que c' est → [/] c' est un peu difficile // parce que → / hhh {%act: inhalation} / c' est → [///] on@oral doit → &eh / la [/] la étudier → // <sinon> ...
*ENT: [<] <hhh {%act: assent}> ///
*ALB: hhh {%act: laugh} ///
*ENT: sinon on@oral [/] <il n' y a rien à faire> ///
*ALB: [<] <mais → > [/] mais en général je crois le &vocabulare [/] le / vocabulaire est / très pareil / avec l' espagnol et → [/] et dans l' anglais il y a → / d' un mínimo@spa → de quarante pour cent de [/] de → [/] de mots qui → [/] <qu' ils>

*ENT: [<] <hhh {%act: assent}> ///
*ALB: ¬ sont utilisés dans la langue // anglaise ///
*ENT: hhh {%act: assent} ///
*ALB: donc → / &eh / hhh {%act: click} / je crois que / pour moi c' est un peu / plus facile que → [/] que [/] que les autres /// <je crois> ///
*ENT: [<] <que pour le reste> /// c' est vrai /// peut-être /// et est ce que → / la façon de faire ici les cours à la [/] à l' école des langues / ça te plait → ? c' est bien faite → ? &eh / est ce que tu changerais quelque chose ? tu mettrais / peut-être → / de différentes choses // quelques idées // à propos de cela → ?
*ALB: hhh {%act: blow} / &mm / hhh {%act: click} / <comme je → > +
*ENT: [<] <je vais rien dire → hein? > ///
*ALB: <hhh {%act: laugh} / comme je → > +
*ENT: [<] <je vais rien dire / hhh {%act: laugh}> ///
*ALB: non mais comme je → [/] je → [/] comme j' ai dit &eh # avant // je crois que → [/] hhh {%act: click} [/] je [/] je [/] je suis très → très optimiste /// parce que comme je t' ai dit / pour moi c' est comme une libération <donc → > ...
*ENT: [<] <hhh {%act: assent}> ///
*ALB: je ne sais pas → // je [/] je crois que / hhh {%act: click} / hhh {%act: blow} # &peut-êt [/] peut-être il sera → nécessaire de [/] de → [/] de nous donner / &mm / plus possibilités de aller à la France / et de → [/] <d' avoir> /
*ENT: [<] <hhh {%act: assent}> ///
*ALB: ¬ quelques programmes → / &eh +
*ENT: d' échanges ?
*ALB: mais → / yo@spa [/] &y [///] cette année pour moi c' est très difficile de le faire // parce que je → [/] je dois être <à la maison / avec mes enfants> ///
*ENT: [<] <hhh {%act: assent}> ///
*ALB: mais → en général je [///] pour moi c' est → [/] c' est parfait // et → ...
*ENT: hhh {%act: assent} ///
*ALB: hhh {%act: click} / parce que → / je [/] je [/] je suis habitué → à → / &eh [/] à une façon de [/] d' apprendre → / hhh {%act: click} / peut-être / très / traditionnelle {%com: ALB laughs while he is speaking} / <et → > ...

*ENT: [<] <hhh {%act: assent}> ///
*ALB: il faut étudier → ///
*ENT: <hhh {%act: assent}> ///
*ALB: [<] <il faut> travailler → // parce que sinon &eh = pour moi c' est → +
*ENT: c' est bien ///
*ALB: c' est bien /// <hhh {%act: assent}> ///
*ENT: [<] <d'accord> /// et → / est ce que tu as visité la France ?
*ALB: hhh {%act: click} / &eh oui → // j' ai → / suis allé → / hhh {%act: click} / quand je → &suvi [/] me suis marié → // je suis → allé → / une semaine {%alt: semane} à Paris ///
*ENT: hhh {%act: assent} ///
*ALB: pendant le → [/] le voyage de noces → ///
*ENT: oui → ///
*ALB: et → [///] parce que j' aime la montagne {%com: ALB pronounces the final 'e' and sounds like Spanish} // et → / hhh {%act: click} / maintenant je ne peux → [/] <je ne peux pas aller → > /
*ENT: [<] <hhh {%act: laugh} / &eh / oui> ///
*ALB: ¬ quand xxx xxx temps mais ... mais je suis allé à Chamonix / &eh / trois fois ///
*ENT: aha! ///
*ALB: parce que / vous savez que à Chamonix <c' est → > +
*ENT: [<] <oui → > /// <hhh {%act: assent}> ///
*ALB: ¬ [<] <avec le Montblanc → > / et les → [/] c' est ... = donc &eh / en été je → [/] je suis allé pendant deux semaines {%alt: semanes} &eh / trois [/] trois ans ///
*ENT: hhh {%act: assent} /// et → / qu' est ce que tu peux me raconter de → [/] de la France → ? &eh / &l → [/] <qu' est ce que tu>
*ALB: [<] <la &éducat> +
*ENT: ¬ penses // de la France ?
*ALB: surtout la → [/] la éducation / que j' ai [/] j' ai trouvé → / là-bas &eh / parce que / hhh {%act: click} / &el [/] &eh [/] la gens c' est → / en France [/] &e [/] en général / &eh / je crois que sont / hhh {%act: inhalation} / plus → polis ? <que → > [/]
*ENT: [<] <hhh {%act: assent}> ///
*ALB: ¬ que ici // je crois &eh / parce que → / yo@spa [/] je &vai [/] je [/] je [/] je → [/] j' ai vu que / pour moi / &eh / dans la montagne / <&eh les → > [/]
*ENT: [<] <hhh {%act: assent}> ///

*ALB: ¬ &eh / les → [/] les alpinistes français → / &eh / sont plus <aimables → > [/]
*ENT: [<] <hhh {%act: assent}> ///
*ALB: ¬ plus → +
*ENT: ils sont <respectueux → > ...
*ALB: [<] <oui> oui ///
*ENT: hhh {%act: assent} ///
*ALB: par &exam [/] par exemple quand &nono [/] quand nous [/] nous restons à les → [/] à les → [/] hhh {%act: click} [/] à l' auberge → de la montagne → // &eh [/] &eh / les français sont &eh [/] &eh [/] être en silence → / <et → / les → > [/]
*ENT: [<] <hhh {%act: laugh}> / <et le reste ... > ///
*ALB: ¬ [<] <les → [/] les gens qui> vient d' Espagne → / <ils boivent → >
*ENT: [<] <ils sont insupportables> ///
*ALB: ¬ et ils parlent très haut → et → [/] et quelquefois ils / &eh / sont allés sans payer ///
*ENT: hhh {%act: assent} ///
*ALB: et → [/] donc → [///]
*ENT: hhh {%act: assent} ///
*ALB: ¬ &eh / c' est mon avis /// mais <je ne sais pas → &eh> +
*ENT: [<] <non mais → oui → > // peut-être /// il y a → [///] non c' est vrai /// il y a toute &u [/] <hhh {%act: click}> [///]
*ALB: [<] <hhh {%act: assent}> ///
*ENT: ¬ c' est une façon différente de concevoir → l' éducation depuis → [/] <depuis>
*ALB: [<] <hhh {%act: assent}> ///
*ENT: ¬ le maternelle // et ça@oral → [/] ça@oral → [/] ça@oral change /// donc c' est vrai / que / il y a beaucoup des différences → {%com: ENT laughs} /// et pour ce qui est → / de la communication avec les français // <est ce que>
*ALB: [<] <hhh {%act: assent}> ///
*ENT: ¬ tu as utilisé le français quant tu étais à → [/] en France ?
*ALB: oui un peu // <&eh / oui> ///
*ENT: [<] <et ça@oral> a été → / difficile facile → ? est ce que / il y avait des problèmes ?
*ALB: hhh {%act: doubt} / non c' est → [/] c' est → [///] je crois c' est [/] c' est plus facile parce que le vocabulare@n que j' utilisais / ce → [/]

c' était vocabulare@n de &them [/] vocabulare@n très → concret et → [/] et → [/] à sujet de la montagne / <de la corde → > //
*ENT: [<] <hhh {%act: assent}> ///
*ALB: ¬ les xxx les glaciars@n et → [///] donc &jis [/] j' utilisais → / &eh / un corpus de [/] de mots très très → +
*ENT: un corpus /// <hhh {%act: laugh}> ///
*ALB: <bueno@spa / &eh> / hhh {%act: laugh} ///
*ENT: non mais c' est bien → ! <hhh {%act: laugh}> ///
*ALB: [<] <&eh> ...
*ENT: <mais moi je fais un>
*ALB: [<] <un corpus de> +
*ENT: ¬ corpus / <donc>
*ALB: [<] <hhh {%act: assent}> ///
*ENT: ¬ c' est pour ça@oral que → / je rigole non? <pas xxx>
*ALB: [<] <donc je crois> [/] je [/] je le trouvais → / <facile> /
*ENT: [<] <hhh {%act: assent}> ///
*ALB: ¬ mais → = hhh {%act: assent} ///
*ENT: <d'accord> ///
*ALB: [<] <et → > ...
*ENT: non mais c' est bien /// hhh {%act: inhalation} +
*ALB: mais je [/] j' aime la → musicalité de la langue <français → > [/]
*ENT: [<] <de la langue> <française> ///
*ALB: ¬ [<] <de la → > / hhh {%act: inhalation} / sutilesse@n → ? <peut-être → > ...
*ENT: [<] <oui> /// hhh {%act: assent} ///
*ALB: <hhh {%act: assent}> ///
*ENT: [<] <et → > / pour ce &qu [///] à part la montagne / &eh / est ce que vous avez des loisirs // différents ? qu' est ce que vous → [/] tu fais pendant le temps libre ? si tu as de temps libre /// hhh {%act: laugh} /// <avec le bébé / avec les enfants> ///
*ALB: <hhh {%act: laugh} / oh! j' aime la> [/] la lecture et → [/] et les [/] les → [/] j' aime &le [/] des activités en plein air et → / la lecture je → [/] j' aime la lecture // je [/] <j' aime>
*ENT: [<] <hhh {%act: assent}> ///
*ALB: ¬ lire → // surtout {%alt: sortout} ///
*ENT: et → quel type de livres ?
*ALB: <hhh {%act: click}> ///
*ENT: [<] <par exemple> // il y a un type concret → ? <xxx>
*ALB: [<] <littérature> anglaise et / normalement / &eh / c' est le → [/] les classiques ///

*ENT: hhh {%act: assent} ///
*ALB: et maintenant j' ai commencé à lire le [/] la littérature <français> +
*ENT: [<] <française> ? <hhh {%act: assent}> ///
*ALB: [<] <et → > [/] et → / hhh {%act: doubt} / ça@oral <m' a [/] m' a ouvert> +
*ENT: [<] <mais en français> <ou → en espagnol> ?
*ALB: [<] <en français → > ...
*ENT: hhh {%act: assent} ///
*ALB: et ça@oral / m' a ouvert → &eh / une nouvelle → +
*ENT: voie de → [///] non? <de → [/] d' exploration> {%com: ENT laughs} ///
*ALB: [<] <oui / oui oui oui> /// et c' est très intéressant comparer les → [/] <hhh {%act: click}> [/]
*ENT: [<] <hhh {%act: assent}> ///
*ALB: ¬ les → [/] les deux / point de vue / xxx / qui sont / dans la littérature <et → > ...
*ENT: [<] <c' est> vrai /// et par exemple / un livre / que → [///] de littérature française ? <que → > +
*ALB: [<] <que j' aime → > / <beaucoup> ?
*ENT: [<] <hhh {%act: assent}> /// hhh {%act: assent} ///
*ALB: ah! Madame Bovary ///
*ENT: Bovary c' est vrai / <c' est → / le livre> ! ///
*ALB: [<] <bien sûr> {%com: ENT laughs} /// mais → +
*ENT: le livre pour tous qui aiment lire // c' est vrai // que c' est le livre /// <hhh {%act: laugh}> ///
*ALB: [<] <je → [/] je vais → > lire / le → [/] oh! [/] La Recherche / Du Temps Perdu → // mais → / je crois que <c' est un>
*ENT: [<] <xxx> +
*ALB: ¬ peu difficile ///
*ENT: oui → // le vocabulaire c' est difficile /// c' est vrai mais → [///] bon // ça@oral dépend /// <il faut commencer → > /
*ALB: [<] <hhh {%act: assent and doubt}> ///
*ENT: ¬ et → voir / qu' est ce que ça@oral donne /// c' est vrai / qu' ils sont des phrases → / très longues mais → / s' il faut → / relire deux fois → // il faut le relire // mais / c' est bien ///
*ALB: j'ai lu / &eh [/] je &j [/] hhh {%act: click} [/] un [/] &eh [/] un autre → / livre / que [/] qui → / s' appelait → / La Première De Corde {%com: the real title is 'Premier de Cordée'} / <de → > /

*ENT: [<] <hhh {%act: assent}> ///
*ALB: ¬ Frison → Roche ? <je crois que c' est>
*ENT: [<] <moi xxx> ///
*ALB: ¬ français → // o@spa &ae [/] alemán@spa je ne sais pas ///
*ENT: moi → je ne sais pas ///
*ALB: mais → ...
*ENT: je connais pas {%oral: 'ne' absent} ///
*ALB: &i [/] il / n' a pas un grand valeur littéraire mais c' est à → / sujet de la montagne et → /
*ENT: hhh {%act: assent} ///
*ALB: ¬ en relation avec la montagne // mais → / hhh {%act: inhalation} [/] mais / hhh {%act: click} / &eh / je crois que je vais lire → / si j' ai → / le temps / <&eh> [/]
*ENT: [<] <hhh {%act: assent}> ///
*ALB: ¬ &e [/] tous les livres d' &an [/] de → [/] en français // <parce que> → /
*ENT: [<] <oui> /// <hhh {%act: assent}> ///
*ALB: ¬ [<] <c' est> très très intéressant // et → ...
*ENT: c' est vrai / non c' est une façon d' apprendre <aussi> ///
*ALB: [<] <hhh {%act: assent}> ///
*ENT: et → / &mm [/] et bon / &eh / on@oral a parlé de la lecture /// &eh / &ci [/] cinéma / est que tu regardes → / quelque chose de cinéma français / ou → tu connais ou → &euh ?
*ALB: je → / hhh {%act: doubt} / &eh [///] le cinéma c' est [/] c' est → [///] ça@oral ne m' intéresse → / beaucoup // <mais → > /
*ENT: [<] <hhh {%act: assent}> ///
*ALB: ¬ le cinéma français → / hhh {%act: inhalation} # aujourd'hui le → [///] il y a → / trois semaines {%alt: semanes} → / xxx un film / &eh / française que [/] que je [/] je l' ai trouvé / très très intéressant /// &eh / Les Amis Du Nort@n {%com: ALB is possibly thinking about 'Bienvenu chez les 'chtis'} ?
*ENT: hhh {%act: assent} ///
*ALB: quizás@spa ?
*ENT: oui ///
*ALB: mais → / &pe [/] je [/] je ne [/] je ne / peux pas parler → / beaucoup de [/] de cinéma <parce que → > +
*ENT: [<] <hhh {%act: laugh}> / non mais → / c' est normal ! <ça@oral dépend → / des intérêts → > +

*ALB: [<] parce que en &compa> [/] en comparaison avec la littérature / je crois qu' il y a une grande différence → ///
*ENT: <hhh {%act: assent}> ///
*ALB: [<] <&eh> / la littérature / on@oral peut trouver → / xxx livres ///
*ENT: hhh {%act: assent} ///
*ALB: mais dans la film [/] dans la [/] le cinéma // c' est très difficile /// pour moi → // de trouver un bon film /// &eh / hhh {%act: blow} / je crois que le cinéma est / hhh {%act: blow} ...
*ENT: hhh {%act: assent} / <bon / ça@oral dépend> ///
*ALB: [<] <je sais que le [/] le> [/] le France [/] la France a → / un cinéma très → +
*ENT: oh! très → caractéristique // <aussi donc> ///
*ALB: [<] <hhh {%act: assent}> /// <&eh / Alain Renaud est → >
*ENT: [<] <mais il xxx ça@oral dépend / il y a → > [///] oui // Truffaut → / Chabrol /// il y a pas mal de → [/] de → / réalisateurs // qui font des films mais c' est normal hein? il y a → / hhh {%act: laugh} [/] il y a / des intérêts pour tout le monde /// et → / hhh {%act: click} / alors → / on@oral a parlé ça@oral → / &mm / est ce que tu utilises / les nouvelles technologies // pour apprendre les langues ? même pour → / tes cours de → [/] d' anglais /// <hhh {%act: laugh}> ///
*ALB: [<] <&eh> / je → [/] j' essaie de [/] de [/] de le utiliser [/] de les utiliser les → [/] hhh {%act: click} = et / maintenant je crois je → [/] je / hhh {%act: inhalation} [/] je utilise → / Youtube / &eh /
*ENT: hhh {%act: assent} ///
*ALB: ¬ too@eng // hhh {%act: doubt} {%com: ALB says this word in English and tries to find it in French} //
*ENT: hhh {%act: laugh} ///
*ALB: ¬ pour l' apprentissage de français → /
*ENT: <aha!> ///
*ALB: ¬ [<] <je le> xxx &eh / avec la chanson → / française et [/] et je crois que / tous les jours j' écoute → / une chanson → / <française> ///
*ENT: [<] <ah!> oui // et → / <quel type de musique> ?
*ALB: [<] <de Moustaki→ > /// j' aime / Moustaki /// <Georges Moustaki> ///
*ENT: [<] <hhh {%act: assent}> ///
*ALB: et → [/] et → / Édith Piaf /// je suis / &eh / <hhh {%act: laugh}> ...

*ENT: [<] <hhh {%act: laugh}> ///
*ALB: c' est très vieille et <très → > +
*ENT: [<] <non mais → > / ce sont de bonnes chansons → // ils ont des paroles → / <vraiment &eh> /
*ALB: [<] <hhh {%act: assent}> ///
*ENT: ¬ bien faites // donc c' est bien ///
*ALB: hhh {%act: assent} ///
*ENT: c' est pas tout le &mon [///] il y a des actuelles que [///] oui → c' est bien mais c' est vrai que → / les anciennes // ils restent toujours ///
*ALB: hhh {%act: inhalation} / mais encore je → [/] je crois que / &eh / dans → &l [/] la chanson française // on@oral trouve / les sentiments → / <plus facilement>
*ENT: [<] <oui> ///
*ALB: ¬ que dans → / la chanson anglaise ///
*ENT: hhh {%act: assent} ///
*ALB: c' est [/] c' est mon avis ///
*ENT: hhh {%act: assent} ///
*ALB: je crois / mais → / hhh {%act: click} [///] oui oui → / &eh / je [/] je utilise les → [/] les chercheurs {&com: moteurs de recherche} / &eh / pour → / &eh / &se [/] chercher l' information de l' Internet // <bien sûr → > ///
*ENT: [<] <hhh {%act: assent}> ///
*ALB: et → = hhh {%act: assent} ///
*ENT: non mais c' est bien bien /// hhh {%act: laugh} /// moi j' utilise aussi donc → / c' est pour ça@oral que je / demande ce type de choses /// et → / &mm / hhh {%act: click} / et bon [/] et bon [///] qu' est ce que je dois → / raconter encore ? {%com: ENT looks for the next question in her papers} ah! moi → / on@oral va dire / que je ne connais pas Txxx /// donc → / &eh / qu' est ce que je pourrais faire / ce soir ? qu' est ce que vous → / pouvez → me dire // à propos [///] qu' est ce que je peux faire ici // à Txxx ? hhh {%act: laugh} {%com: ENT laughs because of the expression of surprise from ALB} ///
*ALB: &eh / hhh {%act: exhalation} # d'accord / je [/] je crois que / hhh {%act: doubt} # &eh / vous deveriez@c → {%com: devriez} / &eh / te promener → &eh [/] &eh [/] autour de Txxx // et &ten [/] hhh {%act: click} [/] te promener et → / hhh {%act: inhalation} / dans la rue /// <mais → > /
*ENT: [<] <hhh {%act: assent}> ///

*ALB: ¬ non seulement / autour de / Zxxx@spa {%com: name of most known square of Txxx} mais j' &ou [///] de mon avis vous devriez → / &eh / te promener → / &eh / hhh {%act: click} / dans la rue qui / sont solitaires ///
*ENT: <aha!> ///
*ALB: [<] <spécialement> quand le [/] &l [/] la nuit est tombée ///
*ENT: hhh {%act: assent} ///
*ALB: avec / la lumière et → [///] peut-être il parait que → / il peut être → / un peu → / hhh {%act: click} / &eh / dangereux mais → [/] <mais> [/]
*ENT: [<] <ah! oui> ///
*ALB: ¬ mais c' est [/] ce n' est → [/] ce n' est pas vrai // <eh?> ///
*ENT: [<] <non mais c' est pas {%oral: 'ne' absent} vrai> {%com: ENT laughs while she is speaking} / <xxx> ...
*ALB: [<] <mais pour> moi c' est la meilleur <&eh> /
*ENT: [<] <hhh {%act: assent}> ///
*ALB: ¬ chose à faire / &eh +
*ENT: ici ///
*ALB: te <promener → > /
*ENT: [<] <un peu> /
*ALB: ¬ <dans la rue solitaire>
*ENT: ¬ [<] <&eh> /
*ALB: ¬ <de Txxx> ///
*ENT: ¬ [<] <se promener> sans → [/] sans un but concret // <non> ?
*ALB: [<] <oui> oui <oui> ///
*ENT: [<] <ça@oral> s' appelle / flâner /// <en français> ///
*ALB: [<] <et si c' est> <possible → / &eh> /
*ENT: [<] <flâner> ///
*ALB: ¬ de [///] je vous conseille de vous perdre // <dans la rue> ///
*ENT: [<] <oui / c' est> / flâner → /// <ça@oral c' est le verbe> ///
*ALB: [<] <xxx xxx> / <flâner@g> ///
*ENT: [<] <flâner> / c' est le verbe <pour → > +
*ALB: [<] <pour> moi c' est la → [/] la meilleure chose qu' on@oral peut faire à Txxx // parce que / avec le tourisme → / hhh {%act: blow} / je crois que l' esprit de [/] de Txxx est &per [/] est perdu → ///
*ENT: <hhh {%act: assent}> ///
*ALB: [<] <s' est [/] s' est> perdu ///
*ENT: oui → /// <il y a → > /

*ALB: [<] <mais → > / <donc il faut sortir de> [/]
*ENT: ¬ [<] <beaucoup de gens / parfois> ///
*ALB: ¬ de [/] de [/] de le centre de ville et de → [/] de aller à les → [/] à les rues ///
*ENT: <xxx> ///
*ALB: [<] <moins> <transitées@n> //
*ENT: [<] <hhh {%act: assent}> ///
*ALB: ¬ je crois que +
*ENT: oublier le typique / <et → > /
*ALB: [<] <hhh {%act: assent}> ///
*ENT: ¬ découvrir un peu → la ville non? le centre ville ///
*ALB: hhh {%act: assent} ///
*ENT: ben <c' est bien> ///
*ALB: [<] <oui oui> ///
*ENT: c' était un bon plan /// <je vais le noter> {%com: ENT laughs} ///
*ALB: [<] <hhh {%act: laugh}> ///
*ENT: &bo [/] bon → {%com: ENT claps her hands} je te remercie // on@oral a fini donc ça@oral sera suffisant ///
*ALB: hhh {%act: assent} ///

APÉNDICE B:

LISTA DE ERRORES FRECUENTES PARA TODOS LOS NIVELES CONTENIDOS EN CORAF

	TIPO DE ERROR	Nº
1	<LING_LEVEL><L><SIG> <GRAM_CAT><NOUN><NOM> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	49
2	<LING_LEVEL><G><CLA> <GRAM_CAT><PREPOSITION><PES> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTRA><HIG>	38
3	<LING_LEVEL><X><MAN> <GRAM_CAT><PRONOUN><POO> <TARGET_MOD><ERROR_DESC><OM> <ETIOLOGY><INTER><IFL1>	31
4	<LING_LEVEL><X><MAN> <GRAM_CAT><ADVERB><ADV> <TARGET_MOD><ERROR_DESC><OM> <ETIOLOGY><INTER><IFL1>	30
5	<LING_LEVEL><G><CLA> <GRAM_CAT><PREPOSITION><PES> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTER><IFL1>	26
6	<LING_LEVEL><L><SIG> <GRAM_CAT><NOUN><NOM> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTER><IFL1>	25
7	<LING_LEVEL><AMB> <GRAM_CAT><ARTICLE><AIN> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	21
8	<LING_LEVEL><X><ORD> <GRAM_CAT><ADVERB><ADV> <TARGET_MOD><ERROR_DESC><WRO> <ETIOLOGY><INTER><IFL1>	18
9	<LING_LEVEL><AMB> <GRAM_CAT><ADJECTIVE><ADJ> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	16
10	<LING_LEVEL><M><MFC> <GRAM_CAT><VERB><VSC> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	14

	TIPO DE ERROR	Nº
11	<LING_LEVEL><G><GEN> <GRAM_CAT><ARTICLE><ADE> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><UNK>	14
12	<LING_LEVEL><G><AUX> <GRAM_CAT><VERB><VCC> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><HIG>	14
13	<LING_LEVEL><G><GEN> <GRAM_CAT><ARTICLE><AIN> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	14
14	<LING_LEVEL><G><GEN> <GRAM_CAT><ARTICLE><ADE> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	13
15	<LING_LEVEL><G><TPS> <GRAM_CAT><VERB><VCC> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	12
16	<LING_LEVEL><G><GEN> <GRAM_CAT><ARTICLE><AIN> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	12
17	<LING_LEVEL><L><SIG> <GRAM_CAT><ADJECTIVE><ADJ> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	12
18	<LING_LEVEL><G><GEN> <GRAM_CAT><ARTICLE><AIN> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><UNK>	11
19	<LING_LEVEL><G><PER> <GRAM_CAT><VERB><VSC> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	10
20	<LING_LEVEL><L><SIG> <GRAM_CAT><ADJECTIVE><ADJ> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTER><IFL1>	10
21	<LING_LEVEL><G><CLA> <GRAM_CAT><PREPOSITION><PES> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTRA><SIM>	10
22	<LING_LEVEL><X><MAN> <GRAM_CAT><PRONOUN><POO> <TARGET_MOD><ERROR_DESC><OM> <ETIOLOGY><INTRA><SIM>	9
23	<LING_LEVEL><G><CLA>	9

	TIPO DE ERROR	Nº
	<GRAM_CAT><PREPOSITION><PES> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><UNK>	
24	<LING_LEVEL><X><MAN> <GRAM_CAT><ARTICLE><AIN> <TARGET_MOD><ERROR_DESC><OM> <ETIOLOGY><INTER><IFL1>	9
25	<LING_LEVEL><G><TPS> <GRAM_CAT><VERB><VSC> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTRA><SIM>	9
26	<LING_LEVEL><L><SIG> <GRAM_CAT><NOUN><NOM> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTRA><SIM>	8
27	<LING_LEVEL><M><MDS> <GRAM_CAT><ADJECTIVE><ADJ> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	8
28	<LING_LEVEL><M><MFC> <GRAM_CAT><VERB><VSC> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><HIG>	8
29	<LING_LEVEL><L><CPV> <GRAM_CAT><PREPOSITION><PES> <TARGET_MOD><ERROR_DESC><AD> <ETIOLOGY><INTER><IFL1>	7
30	<LING_LEVEL><L><CPN> <GRAM_CAT><PREPOSITION><PES> <TARGET_MOD><ERROR_DESC><OM> <ETIOLOGY><INTER><IFL1>	7
31	<LING_LEVEL><G><CLA> <GRAM_CAT><PRONOUN><POR> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTER><IFL1>	7
32	<LING_LEVEL><G><NBR> <GRAM_CAT><ARTICLE><ADE> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	7
33	<LING_LEVEL><X><MAN> <GRAM_CAT><ARTICLE><ADE> <TARGET_MOD><ERROR_DESC><OM> <ETIOLOGY><INTER><IFL1>	6
34	<LING_LEVEL><G><TPS> <GRAM_CAT><VERB><VCC> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTRA><SIM>	6
35	<LING_LEVEL><G><PER> <GRAM_CAT><PRONOUN><POO>	6

	TIPO DE ERROR	Nº
	<TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><UNK>	
36	<LING_LEVEL><G><MOD> <GRAM_CAT><VERB><VSC> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTRA><SIM>	6
37	<LING_LEVEL><G><PER> <GRAM_CAT><PRONOUN><POO> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	6
38	<LING_LEVEL><G><CLA> <GRAM_CAT><PREPOSITION><PES> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTRA><SIM> <ETIOLOGY><INTRA><HIG>	6
39	<LING_LEVEL><G><GEN> <GRAM_CAT><ADJECTIVE><ADJ> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	6
40	<LING_LEVEL><G><GEN> <LING_LEVEL><PHO> <GRAM_CAT><ADJECTIVE><ADJ> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	6
41	<LING_LEVEL><PHO> <GRAM_CAT><ARTICLE><ADE> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><IGN>	6
42	<LING_LEVEL><G><GEN> <GRAM_CAT><ARTICLE><ADE> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	6
43	<LING_LEVEL><L><SIG> <GRAM_CAT><VERB><VSI> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	5
44	<LING_LEVEL><L><SIG> <GRAM_CAT><VERB><VSC> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	5
45	<LING_LEVEL><PHO> <GRAM_CAT><ADJECTIVE><ADJ> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	5
46	<LING_LEVEL><X><MAN> <GRAM_CAT><PREPOSITION><PES> <TARGET_MOD><ERROR_DESC><OM> <ETIOLOGY><INTER><IFL1>	5
47	<LING_LEVEL><G><GEN>	5

	TIPO DE ERROR	Nº
	<LING_LEVEL><PHO> <GRAM_CAT><ADJECTIVE><ADJ> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><UNK>	
48	<LING_LEVEL><G><GEN> <GRAM_CAT><ADJECTIVE><ADJ> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	5
49	<LING_LEVEL><X><RED> <GRAM_CAT><PRONOUN><POO> <TARGET_MOD><ERROR_DESC><AD> <ETIOLOGY><UNK>	5
59	<LING_LEVEL><M><MFC> <GRAM_CAT><VERB><VSC> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	5
60	<LING_LEVEL><L><SIG> <GRAM_CAT><ADVERB><ADV> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTRA><ANA>	5
61	<LING_LEVEL><L><SIG> <GRAM_CAT><ADVERB><ADV> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	5
62	<LING_LEVEL><L><SIG> <GRAM_CAT><VERB><VSC> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL3>	5
63	<LING_LEVEL><G><CLA> <GRAM_CAT><PREPOSITION><PES> <TARGET_MOD><ERROR_DESC><AD> <ETIOLOGY><INTER><IFL1>	5
64	<LING_LEVEL><PHO> <GRAM_CAT><VERB><VSC> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	4
65	<LING_LEVEL><X><MAN> <GRAM_CAT><PREPOSITION><PES> <TARGET_MOD><ERROR_DESC><OM> <ETIOLOGY><INTRA><SIM>	4
66	<LING_LEVEL><G><PER> <GRAM_CAT><PRONOUN><POO> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	4
67	<LING_LEVEL><X><ORD> <GRAM_CAT><PRONOUN><POI> <TARGET_MOD><ERROR_DESC><WRO> <ETIOLOGY><INTER><IFL1>	4
68	<LING_LEVEL><PHO>	4

	TIPO DE ERROR	Nº
	<GRAM_CAT><NOUN><NOM> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	
69	<LING_LEVEL><X><RED> <GRAM_CAT><PRONOUN><POO> <TARGET_MOD><ERROR_DESC><AD> <ETIOLOGY><INTRA><ASC>	4
70	<GRAM_CAT><ARTICLE><ADE> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><IGN>	4
71	<LING_LEVEL><L><SIG> <GRAM_CAT><VERB><VCC> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	4
72	<LING_LEVEL><PHO> <GRAM_CAT><CONJUNCTION><COS> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><IGN>	4
73	<LING_LEVEL><X><ORD> <GRAM_CAT><ADJECTIVE><ADJ> <TARGET_MOD><ERROR_DESC><WRO> <ETIOLOGY><INTER><IFL1>	4
74	<LING_LEVEL><G><TPS> <GRAM_CAT><VERB><VSC> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><UNK>	4
75	<LING_LEVEL><PHO> <GRAM_CAT><PRONOUN><POO> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><IGN>	4
76	<LING_LEVEL><G><CLA> <GRAM_CAT><ADVERB><ADV> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	4
77	<LING_LEVEL><G><CLA> <GRAM_CAT><ARTICLE><APA> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	4
78	<LING_LEVEL><G><TPS> <GRAM_CAT><VERB><VSC> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	4
79	<LING_LEVEL><L><SIG> <GRAM_CAT><NOUN><NOC> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	4
80	<LING_LEVEL><G><CLA> <GRAM_CAT><PRONOUN><POR> <TARGET_MOD><ERROR_DESC><MIS>	4

	TIPO DE ERROR	Nº
	<ETIOLOGY><UNK>	
81	<LING_LEVEL><L><SIG> <GRAM_CAT><VERB><VSI> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTER><IFL1>	4
82	<LING_LEVEL><X><MAN> <GRAM_CAT><VERB><VSC> <TARGET_MOD><ERROR_DESC><OM> <ETIOLOGY><INTRA><SIM>	4
83	<LING_LEVEL><X><MAN> <GRAM_CAT><CONJUNCTION><COS> <TARGET_MOD><ERROR_DESC><OM> <ETIOLOGY><INTER><IFL3>	3
84	<LING_LEVEL><X><MAN> <GRAM_CAT><ARTICLE><APA> <TARGET_MOD><ERROR_DESC><OM> <ETIOLOGY><INTER><IFL1>	3
85	<LING_LEVEL><G><CLA> <GRAM_CAT><ARTICLE><ACO> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><IGN>	3
86	<LING_LEVEL><G><TPS> <GRAM_CAT><VERB><VSC> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><UNK>	3
87	<LING_LEVEL><G><GEN> <GRAM_CAT><ADJECTIVE><ADJ> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><UNK>	3
88	<LING_LEVEL><G><NBR> <GRAM_CAT><VERB><VSC> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	3
89	<LING_LEVEL><L><SIG> <GRAM_CAT><NOUN><NOM> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><UNK>	3
90	<LING_LEVEL><G><CLA> <GRAM_CAT><ARTICLE><ADE> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTRA><HIC>	3
91	<LING_LEVEL><L><SIG> <GRAM_CAT><ADJECTIVE><ADJ> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTRA><HIG>	3
92	<LING_LEVEL><G><NBR> <GRAM_CAT><ARTICLE><AIN> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	3

	TIPO DE ERROR	Nº
93	<LING_LEVEL><PHO> <GRAM_CAT><NOUN><NOM> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><UNK>	3
94	<LING_LEVEL><PHO> <GRAM_CAT><VERB><VSC> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><UNK>	3
95	<LING_LEVEL><X><MAN> <GRAM_CAT><ARTICLE><ADE> <TARGET_MOD><ERROR_DESC><OM> <ETIOLOGY><INTRA><SIM>	3
96	<LING_LEVEL><G><GEN> <GRAM_CAT><DETERMINER><DENDEQ> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><UNK>	3
97	<LING_LEVEL><AMB> <GRAM_CAT><DETERMINER><DEI> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	3
98	<LING_LEVEL><X><ORD> <GRAM_CAT><ADVERB><ADV> <TARGET_MOD><ERROR_DESC><WRO> <ETIOLOGY><UNK>	3
99	<LING_LEVEL><AMB> <GRAM_CAT><ADJECTIVE><ADJ> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	3
100	<LING_LEVEL><G><AUX> <GRAM_CAT><VERB><VCC> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><HIC>	3
101	<LING_LEVEL><X><MAN> <GRAM_CAT><PRONOUN><POR> <TARGET_MOD><ERROR_DESC><OM> <ETIOLOGY><INTRA><SIM>	3
102	<LING_LEVEL><M><MFC> <GRAM_CAT><VERB><VSC> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	3
103	<LING_LEVEL><L><SIG> <GRAM_CAT><NOUN><NOP> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	3
104	<LING_LEVEL><G><NBR> <GRAM_CAT><DETERMINER><DEP> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	3
105	<LING_LEVEL><G><CLA>	3

	TIPO DE ERROR	Nº
	<GRAM_CAT><PRONOUN><POR> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	
106	<LING_LEVEL><G><MOD> <GRAM_CAT><VERB><VSC> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><IGN>	3
107	<LING_LEVEL><M><MFC> <GRAM_CAT><VERB><VSC> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><ANA>	3
108	<LING_LEVEL><G><CLA> <GRAM_CAT><ARTICLE><ACO> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	3
109	<LING_LEVEL><X><MAN> <GRAM_CAT><PRONOUN><POD> <TARGET_MOD><ERROR_DESC><OM> <ETIOLOGY><INTER><IFL1>	3
110	<LING_LEVEL><L><SIG> <GRAM_CAT><NOUN><NOM> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	3
111	<LING_LEVEL><L><SIG> <GRAM_CAT><NOUN><NOM> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><ANA>	3
112	<LING_LEVEL><L><SIG> <GRAM_CAT><VERB><VSC> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTRA><ASC>	3
113	<LING_LEVEL><M><MDS> <GRAM_CAT><ADJECTIVE><ADJ> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><HIG>	2
114	<LING_LEVEL><G><GEN> <GRAM_CAT><ARTICLE><APA> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	2
115	<LING_LEVEL><G><MOD> <GRAM_CAT><VERB><VSC> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	2
116	<LING_LEVEL><L><CPA> <GRAM_CAT><PREPOSITION><PES> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTER><IFL1>	2
117	<LING_LEVEL><AMB> <GRAM_CAT><NOUN><NOM>	2

	TIPO DE ERROR	Nº
	<TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	
118	<LING_LEVEL><G><CLA> <GRAM_CAT><ARTICLE><ACO> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	2
119	<LING_LEVEL><X><RED> <GRAM_CAT><PRONOUN><POO> <TARGET_MOD><ERROR_DESC><AD> <ETIOLOGY><INTER><IFL1>	2
120	<LING_LEVEL><G><PER> <GRAM_CAT><VERB><VSC> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><UNK>	2
121	<LING_LEVEL><G><GEN> <GRAM_CAT><DETERMINER><DEP> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	2
122	<LING_LEVEL><AMB> <GRAM_CAT><VERB><VSC> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	2
123	<LING_LEVEL><G><CLA> <GRAM_CAT><PREPOSITION><PEL> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTER><IFL1>	2
124	<LING_LEVEL><PHO> <GRAM_CAT><VERB><VSC> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	2
125	<LING_LEVEL><L><SIG> <GRAM_CAT><ADVERB><ADV> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTER><IFL1>	2
126	<LING_LEVEL><L><SIG> <GRAM_CAT><NOUN><NOM> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTRA><ANA>	2
127	<LING_LEVEL><AMB> <GRAM_CAT><VERB><VSI> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	2
128	<LING_LEVEL><AMB> <GRAM_CAT><DETERMINER><DENDEQ> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><SIM>	2
129	<LING_LEVEL><G><AUX> <GRAM_CAT><VERB><VCC> <TARGET_MOD><ERROR_DESC><OM>	2

	TIPO DE ERROR	Nº
	<ETIOLOGY><INTRA><SIM>	
130	<LING_LEVEL><L><SIG> <GRAM_CAT><ADJECTIVE><ADJ> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTRA><ANA>	2
131	<LING_LEVEL><AMB> <GRAM_CAT><VERB><VSC> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	2
132	<LING_LEVEL><L><SIG> <GRAM_CAT><ADVERB><ADV> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><ASC>	2
133	<LING_LEVEL><G><TPS> <GRAM_CAT><VERB><VCC> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><UNK>	2
134	<LING_LEVEL><X><MAN> <GRAM_CAT><ARTICLE><AIN> <TARGET_MOD><ERROR_DESC><OM> <ETIOLOGY><INTRA><SIM>	2
135	<LING_LEVEL><X><MAN> <GRAM_CAT><ARTICLE><APA> <TARGET_MOD><ERROR_DESC><OM> <ETIOLOGY><INTRA><SIM>	2
136	<LING_LEVEL><G><GEN> <GRAM_CAT><DETERMINER><DED> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><UNK>	2
137	<LING_LEVEL><G><CLA> <GRAM_CAT><ADVERB><ADV> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><ASC>	2
138	<LING_LEVEL><X><MAN> <GRAM_CAT><PRONOUN><POD> <TARGET_MOD><ERROR_DESC><OM> <ETIOLOGY><INTRA><SIM>	2
139	<LING_LEVEL><X><RED> <GRAM_CAT><ARTICLE><AIN> <TARGET_MOD><ERROR_DESC><AD> <ETIOLOGY><UNK>	2
140	<LING_LEVEL><PHO> <GRAM_CAT><PREPOSITION><PES> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><IGN>	2
141	<LING_LEVEL><G><MOD> <GRAM_CAT><VERB><VSC> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><UNK>	2

	TIPO DE ERROR	Nº
142	<LING_LEVEL><X><MAN> <GRAM_CAT><ADJECTIVE><AJC> <TARGET_MOD><ERROR_DESC><OM> <ETIOLOGY><INTER><IFL1>	2
143	<LING_LEVEL><G><PER> <GRAM_CAT><PRONOUN><POO> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><ASC>	2
144	<LING_LEVEL><G><PER> <GRAM_CAT><VERB><VCC> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><ASC>	2
145	<LING_LEVEL><G><PER> <GRAM_CAT><DETERMINER><DEP> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	2
146	<LING_LEVEL><G><NBR> <GRAM_CAT><NOUN><NOM> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTER><IFL1>	2
147	<LING_LEVEL><G><CLA> <GRAM_CAT><DETERMINER><DEI> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><INTRA><ASC>	2
148	<LING_LEVEL><G><TPS> <GRAM_CAT><VERB><VCC> <TARGET_MOD><ERROR_DESC><MIF> <ETIOLOGY><UNK>	2
149	<LING_LEVEL><G><PER> <GRAM_CAT><PRONOUN><POO> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><UNK>	2
150	<LING_LEVEL><L><SIG> <GRAM_CAT><NOUN><NOM> <TARGET_MOD><ERROR_DESC><MIS> <ETIOLOGY><INTRA><ASC>	2

GLOSARIO DE SIGLAS y ABREVIATURAS

AA: Análisis de Actuación.

AC: Análisis Contrastivo.

ACL: *Association for Corpus Linguistics.*

AD: Análisis del Discurso.

AE: Análisis de Errores.

AELINCO: Asociación Española de Lingüística de Corpus.

ALAO: Aprendizaje de Lenguas Asistido por Ordenador.

ASL: Adquisición de Segundas Lenguas.

BNC: *British National Corpus.*

CALL: *Computer Assisted Language Learning*, o ELAO en España.

CEA: *Computer Aided Error Analysis* o Análisis de Errores con ayuda del ordenador.

CIA: *Contrastive Interlanguage Analysis* o Análisis Contrastivo de la Interlengua.

CMC: *Computer Mediated Communication* o Comunicación Mediatizada por el Ordenador.

COD: Complemento de Objeto Directo

COI: Complemento de Objeto Indirecto

CORAF: Corpus ORal de Aprendientes de Francés.

C-ORAL-ROM: Corpus Oral de Lenguas Romances.

CREA: Corpus de Referencia del Español Actual.

DALF: *Dîplome Approfondi de la Langue Française.*

DDL: *Data Driving Learning* o Aprendizaje a partir de los datos (uso directo de corpus en enseñanza).

DGLFL: *Délegation Génerale à la Langue Française et aux Langues de France.*

DPU: *Corpus for Delayed Pedagogical Use* o corpus para uso pedagógico posterior.

EAO: Enseñanza Asistida por Ordenador.

EFL: *English as Foreign Language* o Inglés como lengua extranjera.

ELAO: Enseñanza de Lenguas Asistido por Ordenador.

ELE: Español Lengua Extranjera.

ELT: *English Language Teaching*, o Metodología de Enseñanza de la lengua inglesa.

EOI: Escuela Oficial de Idiomas.

FLE: *Français Langue Étrangère* o Francés como Lengua Extranjera.

FLLOC: *French Language Learners Oral Corpora.*

FOS: *Français Objectifs Spécificiques*, o francés para usos específicos.

FOU: *Français Objectif Universitaire*, o francés con fines académicos.

FRIDA: *French Interlanguage Database*, corpus de referencia del francés como L2 de la Universidad de Lovaina.

ICLE: *International Corpus of Learner English.*

IL: Interlengua.

IPU: *Corpus for Immediate Pedagogical Use*, o corpus para uso pedagógico inmediato.

LC: Lingüística de Corpus.

LLI-UAM: Laboratorio de Lingüística Informática de la Universidad Autónoma de Madrid.

LN: Lengua Nativa.

L1: Lengua Materna.

L2: Segunda Lengua o lengua meta.

MCER: Marco Común Europeo de Referencia para las lenguas.

MEL: Metodología de Enseñanza de Lenguas.

POS: *Part of Speech*, o parte del enunciado.

PSA: *Present Situation Analysis* o Análisis de la Situación Presente (de la enseñanza de idiomas y de los aprendientes).

TA: Traducción Automática.

TALC: *Teaching and Language Corpora.*

TDA: *Target Situation Analysis* o Análisis de la situación meta (resultados obtenidos y grado de desarrollo de la L2).

TEI: *Text Encoded Iniciative*, o iniciativa de codificación/marcado de textos.

TIC: Tecnologías de la Información y la Comunicación.

TIC(E): Tecnologías de la Información y la Comunicación en la Enseñanza.

UCLM: Universidad de Castilla-La Mancha.

XML: *eXtended Markup Language*. Estándar de lenguaje de marcado para la web y otras aplicaciones informáticas.