ANÁLISIS DE LA PRODUCCIÓN Y DE ERRORES EN UN CORPUS ORAL DE ESPAÑOL COMO LENGUA EXTRANJERA*

Leonardo Campillos Llanos
Universidad Autónoma de Madrid

Resumen: este trabajo resume un estudio realizado sobre la adquisición de la expresión oral de cuarenta universitarios (N = 40) que estudiaban español como lengua extranjera en los niveles A2 y B1 (*MCER*). El método de obtención de datos es una entrevista semiespontánea, tomando cuatro participantes por cada lengua materna (principalmente, portugués, italiano, francés, inglés, neerlandés, alemán, polaco, chino y japonés). Igualmente, se entrevistó a cuatro hablantes nativos de edad similar (grupo de control) para establecer comparaciones acerca de la producción oral y los errores del habla. El estudio parte de la meto-dología de la lingüística de corpus, la didáctica del español, el análisis de errores asistido por ordenador y la investigación en adquisición de segundas lenguas. Uno de los resultados es la recogida de un corpus de interlengua oral de aprendices de español. Dicho banco de datos ha permitido analizar los aspectos formales del habla no nativa (recuento de palabras o categorías más frecuentes), para lo cual se anotó morfológicamente el corpus mediante el analizador informático GRAMPAL. Asimismo, se analizaron los errores de la producción oral en el nivel léxico, gramatical, de pronunciación y pragmático-discursivo. Para este fin se diseñó una tipología de errores y se anotó manualmente el corpus con etiquetas XML.

Palabras clave: adquisición, computacional, corpus, didáctica, entrevista, error, español, lengua materna, lengua extranjera, oral.

Abstract: This paper summarizes the results of a study on the acquisition of the oral expression by forty university learners (N = 40) who studied Spanish as a Foreign Language at A2 (N=20) and B1 (N=20) levels (*CEFR*). The data collection method was a semi-spontaneous interview, and learners were divided into 10 groups (with four participants in each group) according to their mother tongue (L1). Learners' L1 was principally Portuguese, Italian, French, English,

Dutch, German, Polish, Chinese, and Japanese. Likewise, four native speakers of similar age were interviewed with the same procedures (control group) in order to compare spoken language phenomena and errors between native and non-native speakers. The methodology used departs from the following approaches: Corpus linguistics, Didactics of Spanish as a foreign language, Computer-assisted error analysis, and Second language acquisition research. A result from the project is the collection of a Spanish learner language oral corpus for search and analysis purposes. This data bank has been used to examine formal aspects of non-native speech (e.g. word count or most frequent categories). To accomplish this task, the corpus was part-of-speech tagged by means of an automatic tagger for Spanish language (GRAMPAL). Furthermore, the errors were analysed at different linguistic levels: lexis-semantics, grammar, pronunciation, and pragmatics-discourse. For that goal, an error typology was designed and the errors were manually tagged with XML tags.

Keywords: acquisition, computational, corpus, didactics, error, foreign language, interview, mother tongue, oral, Spanish.

1. Introducción

El presente trabajo pretende contribuir al conocimiento del habla de diferentes grupos de aprendices extranjeros. Para ello se ha analizado su lengua en desarrollo (*interlengua*), atendiendo a las categorías empleadas y los errores cometidos durante una entrevista. Las grabaciones fueron transcritas y conforman un banco de datos (*corpus*) de la producción oral de 40 alumnos, cuyas lenguas maternas pertenecen a la familia románica (portugués, italiano y francés), germánica (inglés, neerlandés y alemán), eslava (polaco), sino-tibetana (chino) y a las lenguas de Japón (japonés).

Una de las motivaciones que fundamentan esta investigación es el parco volumen de estudios rigurosos sobre la adquisición del español (especialmente en niveles bajos), lo cual contrasta con el peso demográfico de esta lengua. Este trabajo tiene como objetivo concretar los aspectos

más espinosos de la adquisición del español en los grupos de nivel intermedio-bajo, a fin de mejorar la didáctica o los materiales de enseñanza. Pero también existe un interés tecnológico que se relaciona con la carencia de recursos lingüísticos especializados (como corpus de datos de estudiantes), y que este proyecto pretende paliar aportando un bancos de datos obtenidos de aprendices no nativos.

Las preguntas de nuestra investigación se pueden formular del siguiente modo:

- ¿En qué medida ciertos contenidos del aprendizaje del español causan dificultades a todos los alumnos en la expresión oral?
- ¿Por qué en el habla ciertos grupos de alumnos presentan más problemas que otros en determinados puntos? ¿Cuáles son esos contenidos y en qué grado?

En nuestro análisis abordamos todos los niveles lingüísticos (gramática, léxico, pronunciación y pragmática-discurso), y no ignoramos el hecho de que quizá hemos querido abarcar demasiado empleando únicamente una entrevista oral para obtener los datos –lo cual ha limitado el análisis completo o riguroso de ciertos aspectos–. Hemos optado por un enfoque global, en lugar de limitarnos a estudiar en profundidad un solo punto (p. ej., la adquisición de los clíticos). Es cierto que, actuando así y teniendo en cuenta el volumen de nuestro corpus, la generalización de los resultados puede ser limitada, pero preferimos tratar todas las áreas a fin de aproximarnos a la adquisición del español desde diferentes perspectivas, lo cual que creemos que enriquece el análisis.

Nuestro trabajo se enmarca en la encrucijada de los estudios de adquisición en segundas lenguas, la didáctica del español para extranjeros, el análisis de errores asistido por ordenador (Dagneaux *et al.*, 1998; Barlow, 2005) y la lingüística de corpus. Asimismo, conviene realizar algunas precisiones terminológicas: aunque se distinguen ambos conceptos, hemos usado los términos *español lengua extranjera* y *español como segunda lengua* (en adelante, L2 o LE, respectivamente) para referirnos indistintamente al proceso de adquisición/aprendizaje en el país donde se habla la lengua que se

adquiere. Finalmente, cabe subrayar el carácter aplicado de esta investigación, que ha intentado estrechar la distancia entre la investigación básica y la práctica docente. Uno de los resultados del proyecto ha sido el desarrollo de una interfaz de consulta para el corpus oral, que se piensa destinar a la formación de profesionales de la enseñanza[1].

Este artículo resume los puntos destacados del estudio que se explica con detalle en Campillos Llanos (2012), y se organiza del siguiente modo. La sección 2 aborda los planteamientos teóricos, y seguidamente se presentan la investigación (sección 3) y la metodología (sección 4). La sección 5 explica el análisis formal de la producción (recuento de palabras y unidades léxicas, y uso de categorías) y la sección 6 detalla el análisis de errores (léxicos, gramaticales, de pronunciación y pragmático-discursivos). Por último, se exponen las conclusiones y las posibilidades de trabajo futuro.

2. Planteamientos teóricos

En esta sección, repasamos los paradigmas desde los que se ha estudiado la interlengua, centrándonos después en el análisis de errores asistido por ordenador. Seguidamente apuntamos nociones de la metodología de la lingüística de corpus; y en último lugar, exponemos el marco de investigación con corpus para el análisis de errores.

2.1. Paradigmas de investigación y análisis de la interlengua de los estudiantes

El concepto de la *interlengua* (según el término forjado por L. Selinker, 1972) hace referencia al sistema lingüístico de los aprendices en su proceso de adquisición, y viene siendo referido por otras denominaciones como *dialecto idiosincrásico* (Corder, 1971) o *sistema aproximado* (Nemser, 1971). Las investigaciones sobre la adquisición de L2s o LEs han abordado la interlengua desde diferentes paradigmas, de entre los cuales pueden destacarse los siguientes (véanse más referencias en Larsen-Freeman y Long, 1991[1994]; Mackey y Gass, 2005; o Ellis y Barkhuizen, 2005):

1. Análisis contrastivo.
2. Análisis de la actuación y análisis de la interlengua o de la frecuencia.
3. Análisis del discurso.
4. Análisis de errores.

El Análisis Contrastivo fue desarrollado entre los años 1940 y 1960 por Charles Fries y Robert Lado, y parte de las teorías conductistas de la adquisición de lenguas, que consideran que el lenguaje se adquiere mediante formación de hábitos. La piedra angular de este paradigma es la Hipótesis del Análisis Contrastivo, que defiende que los elementos similares entre la lengua materna (L1, en adelante) y la lengua que se adquiere serán más sencillos de aprender (por transferencia positiva de la L1 a la L2); y a la inversa, los elementos que divergen serán más difíciles de adquirir. Aunque este supuesto ha sido refutado por los datos de las producciones de los alumnos –no siempre los elementos de las lenguas más cercanas tipológicamente son más fáciles de adquirir–, la investigación resultante ha sido fecunda en estudios entre pares de lenguas.

El análisis de la actuación (*Performance Analysis*) estudia las producciones lingüísticas de los aprendices, tanto las correctas como las incorrectas (Ellis y Barkhuizen, 2005: 73). En los años 70 se empezó a aplicar a la lingüística adquisicional, con estudios longitudinales y transversales, y dio lugar, por ejemplo, a estudios sobre el orden de adquisición de los clíticos. Análogamente, el análisis de la interlengua (*Interlanguage Analysis*) o de la frecuencia (*Frequency Analysis*) estudia un contenido gramatical o comunicativo para dar cuenta de su variabilidad en la lengua del aprendiz, desde perspectivas generativas o cognitivas. Una de las críticas que han recibido es la dificultad para describir la variabilidad de la interlengua en diferentes contextos cuando se centran solo en una modalidad de lengua (oral o escrita) (Dussias, 2003: 237).

Por su parte, el *análisis del discurso* considera el lenguaje como *uso*, poniendo en relación los enunciados con el contexto o situación lingüística; en este marco destacan el *análisis*

de la conversación y el *análisis interaccional.* Las críticas a estos enfoques provienen de estudiosos con un perfil más cuantitativo, que los acusan de falta de rigor por su carácter cualitativo. En realidad, la combinación de ambos tipos de métodos enriquece el análisis, como indica Dörnyei (2007).

Finalmente, el Análisis de Errores surge en la década de 1970 como evolución del Análisis contrastivo tras comprobar que sus premisas no tenían el carácter predictivo otorgado. Corder (1971[1991]) establece diferentes pasos en esta metodología:

1. Identificación del error.
2. Descripción y clasificación del error: se tiende a señalar el nivel lingüístico en que se produce y el mecanismo de cambio que lo causa (p. ej. el orden de palabras).
3. Explicación del error: se diagnostica si se debe a interferencia (interlingüístico), si es de desarrollo (intralingüístico) o se explica por otros factores.
4. Valoración de la importancia del error para la comunicación.

Uno de los reproches hacia este enfoque es que no permite describir completamente la competencia lingüística de los aprendices, puesto que se centra solo en lo que se equivocan. Los hablantes pueden *evitar* la producción de un término o estructura y puede resultar opaco esclarecer las dificultades que existen. En cambio, a diferencia de otros paradigmas, el análisis de errores intenta explicar la causa del error, aunque el diagnóstico no resulte fácil de establecer. Por ello, el análisis de errores no ha sido relegado, sino incorporado al *análisis de la producción* (Larsen-Freeman y Long, 1991).

2.2. La lingüística de corpus

Un *corpus* se puede definir como un conjunto de textos auténticos, procesados informáticamente y escogidos según un criterio para un estudio. Rara vez se trata de documentos al azar, sino que reúnen ciertas características con un fin determinado. Así, la etapa de diseño es cru-

cial, pues define cuestiones como las siguientes (Marcos Marín, 1994; McEnery, Xiao y Tono, 2006: 4-ss.; Granger, 2008: 261-263):

- Representatividad: un tamaño de la muestra equilibrado y proporcional.
- Procesamiento informático: metadatos y anotación o etiquetado.
- Disponibilidad: el acceso a los datos (ej. condicionado por tipo de contenidos).

La anotación lingüística del corpus se puede mantener ajena en otro documento independiente (*stand-alone annotation*), o disponerla en cada archivo correspondiente (*embedded annotation* o 'incrustada'), como procedemos en nuestro trabajo.

Otra discusión acerca de la lingüística de corpus es de si se trata de una metodología o de una rama de la Lingüística. Nosotros seguimos el criterio de McEnery *et al.* (2006: 7), y la consideramos una metodología de estudio, no una disciplina como la gramática.

2.3. Tendiendo puentes entre la lingüística de corpus, el análisis de errores y la adquisición del español como L2/LE

Los estudios con corpus de estudiantes (*Learner Corpus Research*) siguen dos métodos:

> Análisis contrastivo de la interlengua (*Contrastive Interlanguage Analysis*): realiza bien un análisis contrastivo entre producciones lingüísticas de nativos y de no nativos –para observar, por ejemplo, el uso excesivo o limitado de una estructura por los aprendices–, o bien entre las producciones de no nativos de diferentes L1 –para estudiar el grado de interferencia lingüística– (Gilquin, 2008). Este método no precisa de etiquetado de errores, aunque en algunos proyectos se realiza la anotación morfológica para comparar el uso de categorías entre nativos y no nativos.
>
> Análisis de errores asistido por ordenador (*Computer-aided error analysis*): se fundamenta en la meto-

dología del análisis de errores, y a las fases establecidas por Corder (1971) se le añade una etapa más, que sería el etiquetado de errores a partir de su descripción y explicación. Mediante las etiquetas para cada incorrección y técnicas de la lingüística de corpus se pueden recuperar las incorrecciones según criterios de búsqueda y extraer frecuencia de errores (Dagneaux *et al.*, 1998).

Cada método tiene sus partidarios (Myles, 2005: 378): el análisis de errores asistido por ordenador interesa a los profesores de idiomas, pero no a los investigadores en adquisición de lenguas, quienes prefieren no estudiar los errores. En nuestra opinión, resulta muy positiva la combinación de metodologías, como intentamos aplicar en este trabajo.

3. Presentación de la investigación

3.1 Diseño del corpus

El objetivo fue construir un corpus de análisis multilingüe con datos de aprendices de nivel plataforma (A2) y umbral (B1) (*MCER*). A fin de lograr resultados comparables, buscamos la misma cifra de informantes por nivel (N = 20). Igual procedimos respecto al número de sujetos de cada L1 (4 alumnos por grupo), logrando 9 grupos homogéneos (más uno heterogéneo). Con todo, no se alcanzó en todos un equilibrio en cuanto al nivel o al sexo del hablante debido a la disponibilidad de los alumnos. En total, se entrevistó a cuarenta (N = 40) estudiantes extranjeros de español, algunos aprendices internacionales del programa ERASMUS, y otros, alumnos de intercambio por otros convenios. La mayoría estudiaban español en el Servicio de Idiomas de la Universidad Autónoma de Madrid, y otros, en los cursos de español de la Universidad Complutense de Madrid. La tabla 1 resume los datos de las grabaciones (incluyendo los nativos): sexo (H, hombre, o M, mujer), lengua materna, nivel (*MCER*), duración de la grabación (respecto a cada informante y al grupo de L1), y turnos de habla por entrevista.

	Archivo	Sexo	L1	Nivel (MCER)	Duración (mm : ss)	Duración grupo L1	Nº de turnos
Lenguas romances	PORMA2	H	Portugués	A2	25:10	1:26:52	524
	PORWA2_1	M		A2	20:09		328
	PORWA2_2	M	Portugués (brasileño)	A2	19:51		462
	PORWB1	M		B1	21:42		496
	ITAMA2	H	Italiano	A2	20:45	1:13:25	540
	ITAWA2	M		A2	13:09		304
	ITAMB1	H		B1	23:16		436
	ITAWB1	M		B1	16:15		280
	FREMA2	H	Francés	A2	24:08	1:23:17	584
	FREWA2	M		A2	20:31		250
	FREMB1	H		B1	21:56		566
	FREWB1	M		B1	16:46		522
Lenguas germánicas	ENGWA2	M	Inglés	A2	15:04	1:20:39	348
	ENGMB1	H		B1	18:44		436
	ENGWB1_1	M		B1	18:02		347
	ENGWB1_2	M		B1	28:49		733
	DUTMA2	H	Neerlandés	A2	18:19	1:16:46	454
	DUTWA2_1	M		A2	17:33		180
	DUTWA2_2	M		A2	23:05		582
	DUTWB1	M		B1	17:49		370
	GERMA2	H	Alemán	A2	18:23	1:13:24	306
	GERWA2	M		A2	19:45		526
	GERWB1_1	M		B1	15:35		284
	GERWB1_2	M		B1	19:41		336
Lenguas eslavas	POLMA2_1	H	Polaco	A2	22:20	1:32:25	510
	POLMA2_2	H		A2	30:28		656
	POLMB1	H		B1	26:46		443
	POLWB1	M		B1	12:51		268
Lenguas sino-tibetanas	CHIWA2_1	M	Chino	A2	18:48	1:17:27	478
	CHIWA2_2	M		A2	18:45		450
	CHIMB1	H		B1	18:56		425
	CHIWB1	M		B1	20:58		449
Lengua japonesa	JAPWA2	M	Japonés	A2	28:52	1:32:41	552
	JAPWB1_1	M		B1	16:28		466
	JAPWB1_2	M		B1	20:59		498
	JAPWB1_3	M		B1	26:22		679
Otras lenguas	FINWA2	M	Finés	A2	20:27	1:19:05	544
	HUNWA2	M	Húngaro	A2	21:28		164
	KORWB1	M	Coreano	B1	21:14		462
	TURWB1	M	Turco	B1	15:56		288
Grupo control	SPAM_1	H	Español	-	18:57	1:22:29	401
	SPAM_2	H		-	26:47		626
	SPAW_1	M		-	16:49		307
	SPAW_2	M		-	19:56		333

Tabla 1. Datos de los participantes en el corpus

3.2. Identificación del error

En la consideración de error, hemos tenido en cuenta las características de la lengua oral. Las estructuras reformuladas las consideramos *fallos* (Corder, 1971) y no las incluimos en el recuento. Por ejemplo, en este enunciado no marcamos el uso de *ser*:

*JHA: la pizzería *es* cerrada [/] *está* &ce [/] cerrada /// (FREWA2)

Una cuestión más compleja es la propia consideración de errores que también son cometidos por hablantes nativos de español en la conversación informal. Por ejemplo, también es frecuente reconocer las siguientes desviaciones en el habla espontánea (debido a la planificación deficiente o improvisada del discurso oral):

- La omisión de ciertas preposiciones ante subordinadas: ej. *me acuerdo* (*de*) *que*
- El uso innecesario de ciertos conectores, especialmente *y, entonces*, y *por eso*.

Al analizar las transcripciones del grupo de control, se comprueba que estos fenómenos también aparecen en el habla nativa, a la que se están expuestos los no nativos. Sucede lo mismo con los siguientes fenómenos (se indica el archivo del grupo de nativos):

- Concordancia de persona errónea (especialmente con el verbo impersonal *haber* o el sustantivo colectivo singular *gente*): *la gente *hablasen* (SPAM_2).
- Concordancia de género o número erróneo (especialmente cuando el modificador no se encuentra cerca del sustantivo): *la plaza que es un poco *nuevo* (SPAM_1).

No obstante, decidimos corregir estas discordancias, estimando que son desviaciones gramaticales. Pero en la consideración de error no solo se ha de tener en cuenta la variación diafásica (registro informal y formal), sino también la variación dialectal, a fin de evitar la hipercorrección en el proceso de etiquetado. Por ejemplo, no son errores ni la pronunciación de *c + e/i* con sonido [s] (*seseo*) ni la de /x/ con sonido [h], dado que estos fenómenos son comunes en las variedades de Sudamérica o el sur de España. Como consecuencia, estos fenómenos dialectales no se marcan como error.

Otras dificultades surgen ante la interpretación de nuestros datos orales, para los cuales no hay más evidencia que el habla grabada –a diferencia del análisis de errores de la producción escrita, donde la forma de las palabras está presente–. De hecho, no solo se registran ciertas

palabras incomprensibles en la grabación –lo cual vuelve borroso el significado del mensaje–, sino que también ciertos errores son ambiguos. Por ejemplo, existen ambigüedades respecto al nivel lingüístico del error, como en el este contexto:

> *AMA: es una → *lengua {%pho: ['lɪŋgwa]}* muy [/] muy <linda> (PORWB1)

A partir del sonido ['lɪŋgwa], podemos tener dos interpretaciones: 1. que es un error formal de nivel léxico (**lingua* en vez de *lengua*); o 2., de pronunciación (*e,* /e/ → [ɪ]). En este caso, la L1 del aprendiz (portugués) no ayuda a despejar la ambigüedad, ya que en dicho idioma la vocal *e* se pronuncia con un sonido más cerrado ([i]), y la palabra que corresponde a *lengua* también se escribe *língua.* En tanto en cuanto no poseemos mayor evidencia de que el alumno conoce la forma de la palabra (a partir de datos escritos), este tipo de ambigüedad queda sin ser resuelta. Con todo, no cabe duda de que la L1 interfiere en este caso específico.

Otros problemas se relacionan con las deficiencias de la metodología del análisis de errores. A veces es imposible diagnosticar la causa de algunas desviaciones porque no podemos conocer el mensaje que pretendía expresar el aprendiz. Veamos el siguiente enunciado en el que una estudiante explica cómo prepara el sushi:

> *MSU: primero / *tiene que comprar* / arroz japonés /// (JAPWB1_2)

En una interpretación, el error estribaría en la concordancia verbal –la alumna usa la tercera persona (*tiene que comprar*) en lugar de la segunda singular, que puede expresar impersonalidad (*tienes que comprar*). En otra interpretación, el error se relacionaría con la omisión de pronombre *se*, si se pretendía usar la pasiva refleja (*se tiene que comprar arroz*). Ambas correcciones son posibles, y el análisis de errores resulta insatisfactorio. No obstante, este tipo de errores idiosincrásicos no parece revelar ninguna estrategia de aprendizaje o de desarrollo, ni ningún fenómeno de

interferencia, así que deberíamos otorgarle una importancia relativa en el conjunto del análisis.

3.3. Tipología de errores

Se ha diseñado una tipología de errores a partir de estudios previos para el inglés (James, 1998; Nicholls, 2003), el francés (Granger, 2003) o el español (Fernández López, 1990; Vázquez, 1999). Una taxonomía ideal de errores debería ser, como explica Granger (2003), consistente, informativa, flexible y reutilizable. Con todo, la mayoría de las clasificaciones encuentran los siguientes problemas (Heift y Schulze, 2007: 102):

- Cobertura: algunos errores pueden quedar sin ser clasificados cuando la taxonomía se aplica a otro corpus, dada la imposibilidad de anticipar nuevos casos.
- Homogeneidad: es difícil una descripción completa y significativa de los errores.
- Características de la clasificación: muchas categorías son dependientes de cada lengua, o no pueden ser distinguidas fácilmente (por ejemplo, en español el verbo *hacer* + expresión de tiempo tiene la misma función que un adverbio de tiempo).

Los siguientes son los criterios para la clasificación de errores en nuestra taxonomía:

- Categoría, elemento fonológico o categoría sintáctica.
- Mecanismo de cambio: el proceso que cause el error:
  - Omisión: ej., artículo: **camarero está*
  - Uso innecesario: ej., **un mi amigo*
  - Mezcla: ej., *novelas de históricos* ⇒*novelas históricas* o *novelas de historia*
  - Orden: ej., colocación de las palabras en el enunciado.
  - Deformación: ej., **higenias* ⇒ *higiene*
  - Falsa selección: ej., **soy satisfecho* ⇒ *estoy satisfecho*
- Nivel lingüístico:
  - Pronunciación: ej., acento léxico: **abanico [a'ßaniko]* ⇒ abaNIco.

    - Léxico-semántico:
        - Errores formales: ej., acuñaciones (**omeleta* ⇒ *tortilla,* de 'omelet').
        - Errores semánticos: ej., confusión entre *ir/venir*.
    - Gramática: ej., conjugación (**tiengo* ⇒ *tengo*).
    - Pragmática-Discurso: coherencia, cohesión y registro.
- Tipo: grupos específicos de error frecuente: ej., *ser/estar* (**soy satisfecho* ⇒ *estoy*).
- Etiología (causa del error):
    - Intralingüística (errores de desarrollo): **escribido* ⇒ *escrito*
    - Interlingüística (interferencia)[2]: ej., calcos (*frutas de mar* ⇒ *marisco*).
    - Desconocida: ej., errores inducidos por la enseñanza o los materiales.

Estas categorías descriptivas pueden ser problemáticas. El grado de precisión al distinguir errores similares (la *granularidad*) puede resultar excesivo y difícil de comprender por los no especialistas. Igualmente, esta propuesta está determinada por nuestro banco de datos y puede necesitar una adaptación para otros análisis. Este problema de *robustez* se debe en parte a la carencia de estándares entre las taxonomías de errores. Por último, respecto a los errores que no se pueden asignar a una categoría, dudamos de si crear una nueva clase para ellos; nuestra decisión evitó el riesgo de generar en exceso subdivisiones para pocos casos.

Errores ambiguos y criterios de clasificación de error

Ante los errores ambiguos, seguimos el principio de mínima corrección (*Minimum Correction Principle*; Köhlmyr, 2003; tomado de Díaz-Negrillo y Valera, 2009). Con un criterio semejante, se procedió en el Cambridge Learner Corpus, en el cual, ante varias interpretaciones, se decidió marcar el error que menos cambio suponga en la estructura original o el que menos severo resulte (Nicholls, 2003). En la clasificación de errores es preciso

observar el contexto discursivo anterior, sobre todo en los errores *concomitantes* (James, 1998: 117), causados por otra desviación en el entorno próximo.

Los criterios más importantes para el marcado y recuento de errores son los siguientes:

- Si concurren varios errores en una palabra o sintagma, contamos todos, incluso si el uso de un elemento no es necesario en la forma correcta; por ejemplo:

  *SOR: una → [/] *un otro* comida (ENGWB1_2)

  Contamos dos errores: la selección de género del sustantivo *comida* (se concuerda en masculino con *un* y *otro*), y el uso innecesario del artículo con el indefinido.
- La tendencia ha sido minimizar el recuento de errores. Por ejemplo, en este caso:

  *DAK: y en otro → / ¿ *olla {%err: ollo}* ? (POLMA2_2)

  contamos un error por asignación de género (en vez de dos, deformación y género).
- En el caso de las palabras erróneas que se repiten en el mismo enunciado, solamente etiquetamos y contabilizamos la incorrección en la última forma; p. ej.:

  *REM: que *era* → [/] que *era* muy bien ('estuvo') (FREMB1)

  En este contexto, corregimos solamente el error en la última forma de *era.*

- Marcamos solo un error en incorrecciones que afectan a varias palabras pero que se deben a la lógica de una misma regla errónea; ej., la asignación de plural a *gente*:

  *AIS: *las* gente siempre *tienen* / tiempo ('la gente … tiene') (ENGWA2)

4. Metodología

El proceso ha consistido en las siguientes fases (figura 1):

1. Compilación del corpus (recogida de entrevistas).
2. Transcripción de las grabaciones.
3. Sincronización del sonido y el texto.
4. Conversión al formato XML.
5. Procesamiento y segmentación de las grabaciones.
6. Etiquetado de errores.
7. Análisis de errores.
8. Anotación morfológica.

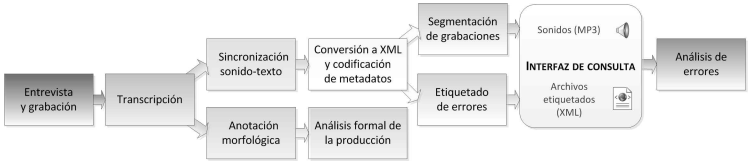


Figura 1. Fases y metodología del proyecto

En los puntos siguientes abordamos dichas fases, excepto el Análisis de errores (§4).

4.1. Compilación del corpus

Los participantes (tanto nativos como no nativos) fueron grabados mientras realizaban una entrevista semiestructurada en la que todos tenían que realizar las mismas tareas. Las técnicas de obtención de datos son similares a las de los exámenes de idiomas:

- Descripción de dos fotografías (en este trabajo, dos tipos de comida).
- Una tarea narrativa a partir de dos historias representadas en viñetas, a fin de que los alumnos usen tiempos de pasado y mecanismos de cohesión discursiva, y para plantearles una pregunta sobre dos actos de habla (una sugerencia y una petición).

En la parte final de la entrevista, los informantes tenían que expresar su opinión acerca de diferentes tópicos (p. ej.,

sobre los cambios en los hábitos de alimentación actual), con el propósito de obtener un habla más espontánea.

4.2. Transcripción de las grabaciones. Marcas de transcripción

Las entrevistas se transcribieron manualmente utilizando el programa Transana© (desarrollado en la universidad de Wisconsin) y fueron revisadas por otro lingüista, a fin de asegurar la consistencia de los contenidos. Las marcas de transcripción son una adaptación del formato CHAT (como se utilizó en el proyecto C-Oral-Rom; Cresti y Moneglia, 2005) y de las convenciones usadas en SPLLOC (Mitchell *et al.*, 2008). En la cabecera de cada archivo se codifican los metadatos de la entrevista, a saber:

- Título y nombre del archivo.
- Datos de los participantes: nombre, sexo, edad, nivel educativo, profesión y origen.
- Fecha, lugar de grabación, situación y tema del discurso.
- Datos sobre la adquisición del español: L1, conocimiento de otras lenguas, nivel, tiempo y contexto de estudio, o tiempo de estancia en país hispanohablante.
- Otras informaciones sobre la entrevista: duración, transcriptor y revisor, etc.

En la transcripción se marcan los siguientes fenómenos (ejemplos de ENGWA2):

- Fenómenos del habla:
  - Pausas: ej., y solo / gramática
  - Actos paralingüísticos/extralingüísticos (risa, tos...): ej. ok hhh {%act: laugh}
  - Solapamientos: ej., *ENT: y <francés> ///
    *AIS: [<] <sí> ///
  - Disfluencias:
    - Repeticiones y reformulaciones: ej. es [/] es muy diferente
    - Apoyos vocálicos o titubeos: ej., &ah &mm / me gusta / &ah leer
    - Palabras fragmentadas: &ma [/] mucho // sí
- Fenómenos del proceso de adquisición de lenguas extranjeras:

- Errores de pronunciación: ej. son más / como / tranquilo {%pho: [tran'kwilo]}
- Errores de deformación: ej., &meji [/] mejillones {%err: melijones} ///
- Acuñaciones o neologismos: ej., el cabarero@n está / enfadada

Los datos personales del aprendiz (como el nombre o el apellido) fueron anonimizados para proteger la identidad de los participantes.

4.3. Sincronización del sonido y el texto. Procesamiento y conversión a XML

La sincronización del texto transcrito con cada segmento de tiempo es una tarea manual realizada con el programa Transana©. El proceso consiste en ir escuchando la grabación e insertar cada marca temporal de final de fragmento mediante una combinación de teclas. Una vez añadidos los códigos de tiempo en las transcripciones, los archivos se procesan y convierten al formato XML. Los archivos procesados incluyen los códigos de tiempo correspondientes a cada fragmento, y dichas marcas se usan para segmentar las grabaciones mediante un programa que automatiza el proceso.

4.4. Etiquetado de errores

El etiquetado se realiza mediante el lenguaje de marcado extensible (eXtensible Markup Language, XML), un formato flexible para el diseño personalizado de etiquetas que ya se ha empleado en proyectos de lingüística, digitalización de textos o corpus de errores (respecto a este tema, véase Izumi *et al.*, 2005; Lüdeling *et al.*, 2005; o Granger *et al.*, 2007). Los elementos de cada fichero XML son los siguientes:

- Metadatos de la entrevista: corresponden a la cabecera de cada transcripción.
- Transcripción: incluye los códigos de tiempo de cada enunciado y turno.
- Etiquetas de errores.

El objetivo de este diseño de datos es permitir la recuperación automática de los fragmentos de texto y sonido donde aparecen errores específicos (según se han definido en las etiquetas de error). En las secciones siguientes explicaremos el conjunto de etiquetas usado para codificar las transcripciones y para marcar y describir los errores.

4.4.1. Etiquetas XML para los contenidos y los metadatos de la transcripción

El diseño de las etiquetas se basa en las usadas en C-Oral-Rom (Cresti y Moneglia, 2005), a las que se añadieron otras específicas que recogen datos sobre la adquisición del español. Los metadatos de la entrevista (que se registran en la cabecera) codifican las variables consideradas en el análisis de errores, entre las que se suele distinguir las *de la tarea* y las *del estudiante* (Barlow, 2005: 337; Granger, 2008: 263-ss.):

- Variables de la tarea (o del contexto): p. ej. condiciones de realización de la prueba.
- Variables del estudiante: p. ej., origen, nivel de competencia (*MCER*), L1, etc.

Algunos de los descriptores (como el nivel educativo o de competencia en español) deben ser tomados con cautela a causa de variabilidad que puede registrar cada categoría (p. ej., debido a la dificultades en alcanzar clasificaciones similares de nivel entre diferentes instituciones o países; Barlow, 2005: 337).

Por lo que respecta a la transcripción, el formato de etiquetas correspondiente a cada unidad de sincronización entre texto y sonido se ha diseñado para codificar interrupciones del otro hablante, o solapamientos de habla, que son muy frecuentes en la conversación espontánea.

4.4.2. Etiquetas XML para los errores

Las etiquetas para el marcado de errores se han diseñado teniendo en consideración los etiquetarios de proyectos anteriores (Granger, 2003; Izumi *et al.*, 2004; Nicholls, 2003; Lüde-

ling *et al.*, 2005; véase un resumen en Díaz-Negrillo y Fernández, 2006). En los archivos XML, se identifica el error en el turno de habla del aprendiz, se describe y se explica mediante los criterios expuestos en la tipología de errores. Un conjunto de etiquetas permite marcar también los errores ambiguos, incluyendo dos interpretaciones diferentes. El diseño de las etiquetas permite marcar errores dentro de otros, en una estructura recursiva para identificar errores que coaparecen en una misma palabra o sintagma; por ejemplo, un error de pronunciación dentro de otra estructura incorrecta.

4.5. Anotación morfológica

Para el estudio cuantitativo de la producción y uso de categorías por los aprendices es necesario contar con corpus anotados morfológicamente. Creemos que la anotación morfosintáctica aporta más ventajas que inconvenientes para el análisis (véase discusión en McEnery, Xiao y Tono, 2006: 29-ss.). El proceso de anotación es semiautomático: en primer lugar, se usa un analizador morfológico informático (GRAMPAL; Moreno Sandoval y Guirao, 2006) que asigna categorías a cada palabra. En segundo lugar, se revisa la anotación mediante una interfaz para la revisión asistida por ordenador (figura 2)[3].

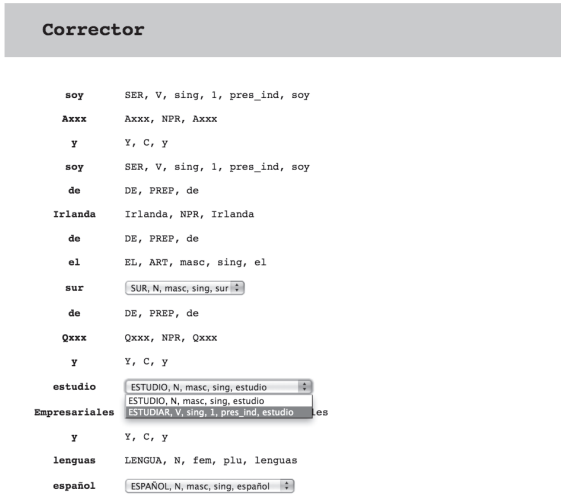


Figura 2. Interfaz de GRAMPAL para la revisión de la anotación morfológica

Una muestra del resultado final de la anotación en XML se presenta en la figura 3.

```
<w cat="V" lem="SER" num="sing" per="1" tie="pres_ind" for="soy" id="1-1-3">soy</w>
<w cat="NPR" lem="Axxx" for="Axxx" id="1-1-4">Axxx</w>
<w cat="C" lem="Y" for="y" id="1-1-5">y</w>
<w cat="V" lem="SER" num="sing" per="1" tie="pres_ind" for="soy" id="1-1-6">soy</w>
<w cat="PREP" lem="DE" for="de" id="1-1-7">de</w>
<w cat="NPR" lem="Irlanda" for="Irlanda" id="1-1-8">Irlanda</w>
<a for="del"> del
<w cat="PREP" lem="DE" for="de" id="1-1-9"/>
<w cat="ART" lem="EL" gen="masc" num="sing" for="el" id="1-1-10"/>
</a>
<w cat="N" lem="SUR" gen="masc" num="sing" for="sur" id="1-1-11">sur</w>
```

Figura 3. Muestra de un fragmento del archivo ENGWA2 anotado con GRAMPAL

5. Análisis cualitativo y formal

5.1. Recuento de palabras y unidades léxicas por grupos de L1

Hemos realizado un recuento en bruto del número de palabras, y de una manera menos mecánica, de unidades léxicas, según los criterios que define Fernández López (1990):

- Palabras simples: *patata, ella, compramos, una, verdaderamente…* En el caso de las contracciones (*al* y *del*), contamos separadamente cada unidad.
- Palabras compuestas que forman una unidad de significado: cuantificadores (*un poco*), sustantivos (*colegio mayor*) o verbos (formas compuestas: *he estudiado*; perífrasis: *va a comer*; pasiva: *son obligados*; y verbos pronominales: *levantarse*).
- Locuciones verbales (*tomar en cuenta*), adjetivales (*de toda la vida*), preposicionales (*alrededor de*), adverbiales (*al final*) o conjuntivas (*para que*).
- Marcadores discursivos (*es decir, por ejemplo*).
- Fórmulas oracionales o rutinarias: *de nada* (para responder a un agradecimiento).
- Nombres propios (*Sierra Morena*), materias académicas (*Filología Clásica*), eventos históricos (*Segunda Guerra Mundial*) o títulos de películas (*El perro*).

Cuando una parte de la unidad léxica se repite (debido a las disfluencias de la oralidad: ej. *a [/] a veces*), consideramos la repetición dentro de una misma unidad. Exceptuamos los

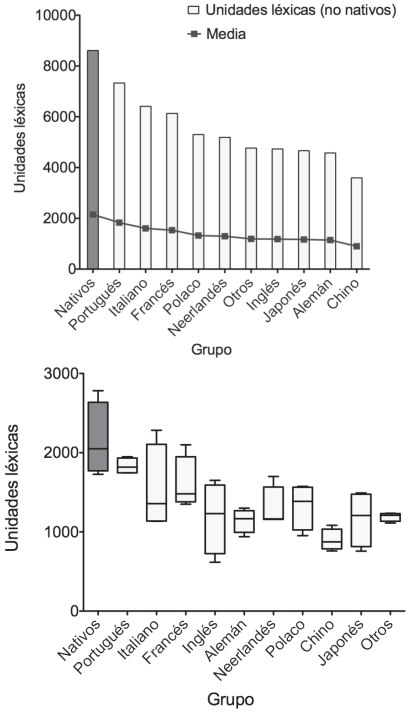
casos en que una parte aparece en otro turno porque el entrevistador se la proporciona. En el caso de los calcos formales (p. ej., **frutas de mar,* 'marisco'), preferimos contar conjuntamente todas las palabras, dado que forman una unidad semántica para expresar un concepto en la L1 de los aprendices (aunque no en español).

La tabla 2 y los gráficos 4 y 5 muestran el recuento de palabras en bruto y por unidades léxicas, así como la media y la desviación estándar (D) de cada grupo.

L1	Archivo	Recuentos en bruto				Unidades léxicas			
		Archivo	Grupo	Media	D	Archivo	Grupo	Media	D
Portugués	PORMA2	1860	7712	1928,00	103,76	1748	7329	1832,25	102,21
	PORWA2_1	1845				1745			
	PORWA2_2	1935				1887			
	PORWB1	2072				1949			
Italiano	ITAMA2	1710	6603	1650,75	569,69	1572	6133	1533,25	539,92
	ITAWA2	1200				1135			
	ITAMB1	2434				2283			
	ITAWB1	1259				1143			
Francés	FREMA2	1538	6760	1690,00	347,21	1467	6413	1603,25	337,13
	FREWA2	1563				1496			
	FREMB1	2206				2100			
	FREWB1	1453				1350			
Inglés	ENGWA2	1115	5009	1252,25	483,16	1049	4731	1182,75	451,18
	ENGMB1	1767				1651			
	ENGWB1_1	648				617			
	ENGWB1_2	1479				1414			
Neerlandés	DUTMA2	1195	5473	1368,25	279,22	1160	5186	1296,50	269,03
	DUTWA2_1	1230				1158			
	DUTWA2_2	1785				1700			
	DUTWB1	1263				1168			
Alemán	GERMA2	1241	4800	1200,00	148,38	1159	4573	1143,25	149,53
	GERWA2	1194				1174			
	GERWB1_1	1004				940			
	GERWB1_2	1361				1300			
Polaco	POLMA2_1	1280	5546	1386,50	310,29	1241	5300	1325,00	288,89
	POLMA2_2	1679				1575			
	POLMB1	1589				1531			
	POLWB1	998				953			
Chino	CHIWA2_1	895	3731	932,75	139,25	862	3592	898,00	135,50
	CHIWA2_2	1118				1084			
	CHIMB1	783				760			
	CHIWB1	935				886			
Japonés	JAPWA2	783	4953	1238,25	397,44	757	4664	1166,00	354,03
	JAPWB1_1	1028				985			
	JAPWB1_2	1549				1430			
	JAPWB1_3	1593				1492			
Otros	FINWA2	1289	4990	1247,00	66,66	1236	4767	1191,75	54,80
	HUNWA2	1148				1112			
	KORWB1	1273				1215			
	TURWB1	1280				1204			
Total		55577		1389,425	399,48	52688		1317,20	374,16
Español (grupo de control)	SPAM_1	2363	9389	2347,25	514,14	2195	8610	2152,50	462,47
	SPAM_2	3054				2783			
	SPAW_1	1859				1727			
	SPAW_2	2113				1905			

Tabla 2. Número de palabras (recuento en bruto) y de unidades léxicas por grupos de L1

En nuestro corpus, la producción más baja se recoge los grupos chino (con diferencia), alemán y japonés (por este orden). En cambio, los lusófonos, franceses e italianos presentaron la producción más alta. Estos resultados podrían asociarse a causas interlingüísticas (la cercanía o lejanía entre la L1 y el español explicaría sus dificultades de expresión), pero también a factores socioculturales (su participación activa o pasiva).



Figuras 4 y 5. Unidades léxicas; frecuencia absoluta y media de grupo, y representación de desviación

5.2. Categorías más utilizadas

Las listas de frecuencia de categorías se obtuvieron procesando informáticamente los archivos anotados con GRAMPAL y haciendo uso de la herramienta SketchEngine (Kilgarriff *et al.*, 2004). La tabla 3 desglosa el número de unidades léxicas por categoría en cada grupo (el total

de unidades se ordena de mayor a menor frecuencia absoluta).

Cat / L1	POR	ITA	FRE	ING	NER	ALE	POL	CHI	JAP	OTH	TOTAL	NATIVOS
Adverbios	1043	832	1093	884	792	898	884	668	843	889	8826	935
Verbos	1336	1070	1041	770	875	709	796	566	761	861	8785	1631
Sustantivos	1112	894	907	685	733	661	839	655	831	703	8020	1183
Conjunciones	771	649	718	516	646	468	550	329	539	601	5787	983
Preposiciones	691	623	595	378	470	422	604	284	347	381	4795	764
Artículos	572	554	550	359	441	364	318	270	221	252	3901	685
Adjetivos	301	219	232	134	194	170	253	167	219	212	2101	298
Pronombres	319	323	174	154	193	153	185	121	149	185	1956	592
Marc. discurs.	201	197	352	191	155	133	109	72	163	127	1700	582
Indefinidos[4]	268	212	150	151	199	104	217	115	126	160	1702	289
Nom. propios	152	92	122	106	85	86	115	96	128	103	1085	65
Numerales	71	76	97	41	72	60	52	55	68	51	643	91
Interjecciones	71	15	61	99	54	81	35	55	84	45	600	32
Relativos	94	129	98	33	78	23	44	16	7	28	550	208
Desconocidas	79	40	19	58	53	75	117	40	27	25	533	49
Demostrativos	97	70	36	23	46	43	68	22	44	37	486	98
Interrogativos	57	25	54	42	46	35	58	25	38	35	415	40
Posesivos	47	59	54	44	25	26	33	23	42	49	402	49
Fórmulas	35	39	53	57	23	56	17	1	24	14	319	26
Auxiliares[5]	12	15	7	6	6	6	6	12	3	9	82	10
Total	7329	6133	6413	4731	5186	4573	5300	3592	4664	4767	52688	8610

Tabla 3. Distribución de unidades léxicas por categorías en cada grupo de L1

Se puede observar que el uso de ciertas categorías es muy pobre en ciertos grupos de alumnos, lo cual revelaría problemas interlingüísticos por ausencia en su L1, a saber:

- Artículos entre polacos, chinos, japoneses y el grupo heterogéneo.
- Relativos y preposiciones entre japoneses y chinos.
- Conjunciones, marcadores discursivos y fórmulas orales en el grupo chino.

6. Análisis de errores

En este trabajo se ha distinguido entre errores no ambiguos y ambiguos. Por lo que se refiere a estos últimos, hemos registrado un total de 304 ambigüedades, solo el 4,45% de todos los errores (6838). Esto quiere decir que el 95,55% del análisis de errores realizado no representa una grave ambigüedad, y que la proporción entre ambiguos y no ambiguos que hemos encontrado es de un error ambiguo cada más de 22 aproximadamente. En la mayoría de las ocasiones la ambigüedad estriba en el nivel lingüístico del error, siendo tan solo unos pocos (16) ambiguos respecto a otros criterios, aunque resulta clara su adscripción a un determinado nivel. Los tipos de ambigüedades registradas han sido las siguientes (figura 6):

- Gramática-léxico (4,61%): p. ej., *me gusta // ¿ película ?* (gramática, uso del plural: 'las películas'; léxico: 'el cine').
- Gramática-pronunciación (14,47%): p. ej., *llegue* ['ʝeɣe] (acento o error verbal).
- Léxico-pronunciación (75,66%): ej., *lengua* ['liŋgwa] (préstamo o fonética).
- Pronunciación-pragmática (0,33%): p. ej. *la camarera* [kama'ɾerə] *responda que…* (pronunciación de vocal /a/, o problema de comprensión de la historia).
- Gramática-pragmática (0,33%): p. ej., [Refiriéndose en segunda persona en un contexto informal] *¿Quiere ir a otro restaurante?* (gramática, uso de la persona verbal correcta; pragmática, uso de un registro no formal).

- Ambigüedades dentro del nivel gramatical (3,62%): *cambio la hora* (omisión de pronombre: 'se cambia'; concordancia de persona: 'cambias').
- Ambigüedades en el nivel léxico (0,99%): *tomas un *plazo* (selección léxica de verbo: 'necesitas un plazo [para pedirlo]' o relación semántica: 'pides una plaza').

El mayor número de ambigüedades se registró, con diferencia, entre nivel léxico y de pronunciación; y en segundo lugar, entre gramática y pronunciación (figura 6).

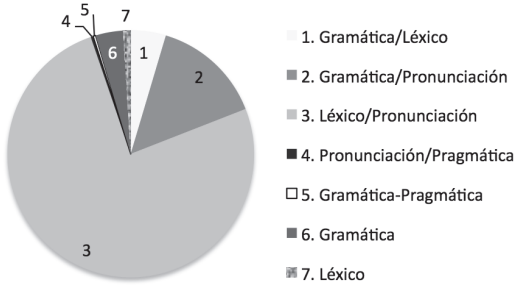


Figura 6. Frecuencia de cada tipo de error ambiguo

La tabla 5 y la figura 7 resumen las cifras de error de cada tipo y nivel lingüístico.

Errores			*Total*
	Nivel lingüistico	*Total*	*(%)*
No ambiguos respecto al nivel	Pronunciación	970	14,19%
	Gramática	3324	48,61%
	Léxico-semántico	2008	29,37%
	Pragmática-Discurso	245	3,58%
	No asignado	3	0,04%
Ambiguos		288	4,21%
Total		6838	

Tabla 4. Frecuencia relativa (%) de errores no ambiguos y ambiguos por nivel lingüístico

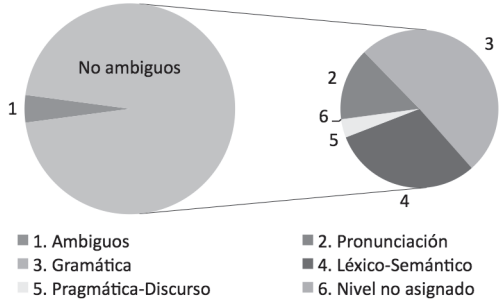


Figura 7. Errores ambiguos y no ambiguos (estos, desglosados por niveles lingüísticos).

Los errores de gramática son los más frecuentes en el habla de los alumnos, seguidos de los errores léxico-semánticos. Estos resultados coinciden con el análisis de errores escritos realizado por Fernández López (1990). A pesar de estas cifras, no todos los errores de gramática impiden la comunicación; p. ej., muchos errores de concordancia (p. ej., *las cosas *españoles* -> españolas), mientras que ciertos errores léxicos son realmente ambiguos respecto a la información expresada. Cabe señalar que no clasificamos en ningún nivel 3 errores que parecen más bien desviaciones idiosincrásicas o de adecuación (un hablante nativo no usaría tales construcciones).

5.1. Recuento global

El total de errores marcados es 6838, incluyendo errores no ambiguos (6534, el 95,55% del total de errores) y ambiguos (304, el 4,45%), con una media de 171 errores por estudiante prácticamente (tabla 6). Como se esperaba de los resultados obtenidos en otros análisis a partir de textos escritos (Fernández López, 1997 [1990]: 65), la tasa de error en el nivel A2 cae notablemente en el nivel B1. Las tablas 6 y 7 y el gráfico 8 muestran la frecuencia absoluta y relativa (expresada en %) de los errores (no ambiguos y ambiguos) respecto al nivel de competencia (*MCER*).

TOTAL DE ERRORES (VALORES ABSOLUTOS)						
Nivel	*No ambiguos*		*Ambiguos*		*Total*	
	Err.	*Media*	*Err.*	*Media*	*Err.*	*Media*
A2	3826	191,30	218	10,90	4044	202,50
B1	2708	135,40	86	4,30	2794	139,25
Total	6534	163,35	304	7,60	6838	170,63

Tabla 5. Frecuencia de errores no ambiguos y ambiguos de acuerdo al nivel del estudiante (MCER).

TOTAL DE ERRORES (PORCENTAJES)				
Nivel MCER	No ambiguos		Ambiguos	
	Fre. relativa	% sobre un. léxica	Frec. relativa	% sobre un. léxica
A2	56,95% (sobre 6838)	14,54% (26317)	3,19% (sobre 6838)	0,83% (26317)
B1	39,60% "	10,27% (26371)	1,26% "	0,33% (26371)
Total	95,35% "	12,40% (52688)	4,45% "	0,58% (52688)

Tabla 6. Frecuencia relativa de errores y proporción por unidad léxica de cada nivel (expresadas en %).

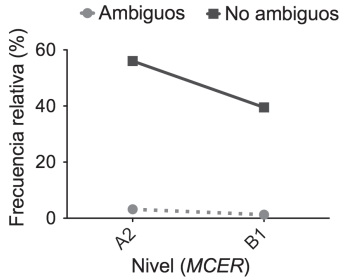


Figura 8. Frecuencia relativa de errores por nivel (% sobre el total de errores cometidos)

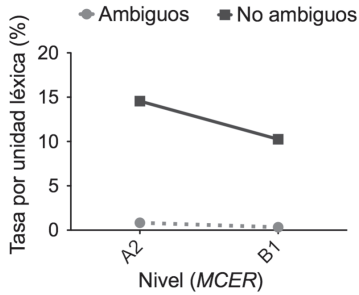


Figura 9. Tasa de error por unidad léxica por nivel (% sobre el total de unidades léxicas de cada nivel)

Los siguientes son los 20 errores más frecuentes según el corpus y nuestra corrección:

Tipo de error	*Nivel lingüístico*	*No ambiguos*	*% del total de errores (6838)*
1. Extranjerismos	Léxico	875	12,80%
2. Artículo	Gramática	634	9,27%
3. Preposiciones	Gramática	541	7,91%
4. Formación no atestiguada	Léxico	359	5,25%
5. Pronombres personales	Gramática	302	4,42%
6. Género	Léxico	272	3,98%
7. Estructura de la oración	Gramática	248	3,63%
8. Concordancia de persona	Gramática	202	2,95%
9. Relación semántica	Léxico	196	2,87%
10. Subordinación	Gramática	187	2,73%
11. Concordancia de número	Gramática	184	2,69%
12. Concordancia de género	Gramática	183	2,68%
13. Tiempos de pasado	Gramática	152	2,22%
14. Conjugación	Gramática	136	1,99%
15. Marcadores discursivos	Pragmático-discursivo	133	1,95%
16. *Ser/estar*	Gramática	122	1,78%
17. Oclusivas sordas/sonoras	Pronunciación	122	1,78%
18. Acento	Pronunciación	118	1,73%
19. Cambios de punto/modo art.	Pronunciación	72	1,05%
20. Pérdida de consonantes	Pronunciación	71	1,04%

En ciertos puntos las cifras realmente reflejan una frecuencia de error acusada solo en ciertos alumnos: los extranjerismos se concentran en el grupo lusófono; los artículos, en los japoneses, chinos, polacos y el grupo heterogéneo; los cambios de acento, sobre todo en el grupo francés; y la no distinción entre oclusivas sordas y sonoras, en los chinos. Los errores más frecuentes y generalizados se relacionan con las preposiciones, las formaciones no atestiguadas, los pronombres personales, la asignación de género (excepto en el grupo italiano), la estructura de la oración, la subordinación, las concordancias y los tiempos de pasado. Por niveles, se pueden apuntar estas tendencias:

- En léxico, sobresalen los errores formales: en A2, deformaciones y extran-jerismos (existe gran transferencia entre nativos de lenguas románicas, sobre todo lusófonos); y aunque en B1 disminuyen su frecuencia, persisten otros errores por relación semántica y asignación de género.
- En gramática, además de los puntos comentados, destaca la alta cifra de incorrecciones por omisión y orden de palabras, las discordancias y el abuso del presente sobre otros tiempos, que se explicaría por las características de la oralidad.
- En pronunciación, el problema más generalizado es el sonido [r], aunque cada grupo tiende a transferir rasgos: p. ej. chinos y japoneses no distinguen /ɾ/ y /l/, o los nativos de lenguas germánicas cometen más errores por reducción vocálica.
- En pragmática-discurso destacan los errores de cohesión (p. ej. marcadores discursivos), aunque se registra gran variabilidad individual entre los alumnos.

Respecto a la etiología de los errores, cabe decir que los resultados han de tomarse con prudencia, ya que en muchos contextos presuponen conocer el mensaje pretendido por el hablante, dando rienda suelta a nuestras (mejores o peores) capacidades interpretativas. De hecho, cuando consultamos la literatura de investigación sobre el tema, constatamos las diferentes conclusiones a las que llegan distintos investigadores. Por ejemplo, según Dulay

y Burt (1976), los errores de interferencia no llegan al 5%, mientras que Nickel (1981) los cifra en una media del 60% (otros lingüistas calculan incluso hasta un 80%); en realidad, la definición de lo que se considera *errores de interferencia* condiciona los resultados (véase un resumen en Lott, 1983). Con todo, estimamos que al menos el 49,21% de las incorrecciones se debe a causas interlingüísticas (tabla 8 y figura 10), lo cual parece estar en la línea de los resultados de Lott (1983) para el inglés escrito, o de Valverde Mateos (2012) respecto al francés oral.

Etiología	*Errores*			
	No ambiguos	*Ambiguos*	*Total*	*F. r. (%)*
1. Interlingüísticos	3274	91	3365	49,21%
2. Intralingüísticos	1534	27	1561	22,83%
3. Desconocida	1726	46	1772	25,91%
4. Varias causas	-	140	140	2,05%
Total	6534	304	6838	

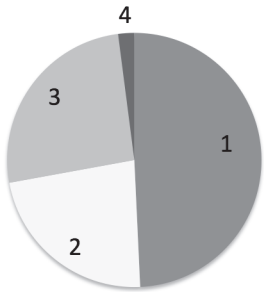


Tabla 7 y Figura 10. Etiología estimada de los errores (frecuencia absoluta y relativa en %)

La distribución por niveles revela un descenso de los interlingüísticos de A2 a B1, y al contrario, un leve aumento de intralingüísticos, como apuntan otros trabajos (Vázquez, 1999) (tabla 9 y gráfico 11; la frecuencia relativa se mide respectivamente sobre el total de errores de cada nivel: por una parte, los no ambiguos; por otra, los ambiguos; y el total, sobre ambos; la cifra de cada uno se indica entre paréntesis). La alta tasa de errores de etiología desconocida se debe a nuestro criterio prudente de marcado.

Errores		*Etiología*									
		Interlingüísticos		*Intralingüísticos*		*Desconocida*		*Varias causas*[4]			
		Frec	*Frec. rel. (%)*	*Frec*	*Frec. rel. (%)*	*Frec*	*Frec. rel. (%)*	*Frec*	*Frec. rel. (%)*
A2	No ambiguos	2024	52,92% (3825)	809	21,15% (3825)	988	25,93% (3825)	-	-
	Ambiguos	85	38,99% (218)	9	4,13% (218)	45	20,64% (218)	79	36,24% (218)
	Total	2109	52,16% (4043)	818	20,35% (4043)	1033	25,65% (4043)	79	1,95% (4043)
B1	No ambiguos	1249	46,11% (2709)	726	26,80% (2709)	725	27,09% (2709)	-	-
	Ambiguos	6	6,98% (86)	18	20,93% (86)	1	1,16% (86)	61	70,93% (86)
	Total	1255	44,90% (2795)	744	26,62% (2795)	726	26,30% (2795)	61	2,18% (2795)

Tabla 8. Distribución de errores por nivel según la etiología (frecuencia absoluta y relativa en %)

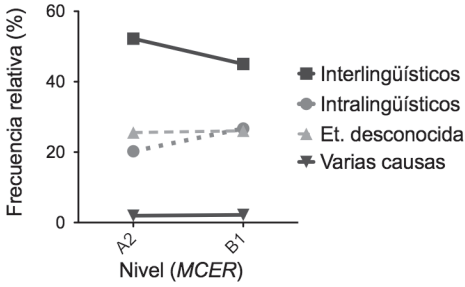


Figura 11. Etiología de errores (no ambiguos y ambiguos) por nivel (frecuencia relativa en %)

6. Conclusiones

Esta investigación ha analizado la interlengua oral de cuarenta estudiantes de español. Los resultados del análisis formal de la producción en nuestro corpus son los siguientes:

- En el recuento de unidades, apenas hay diferencia entre A2 y B1, pero es más baja en chinos, alemanes y japoneses, y más alta en lusófonos, franceses e italianos.
- En grupos específicos de L1 hay una producción más pobre de ciertas categorías, quizá por causas interlingüísticas: artículos (polacos, chinos y japoneses), relativos y preposiciones (japoneses) y conjunciones y marcadores discursivos (chinos).

El análisis de errores de nuestro corpus reveló las siguientes dificultades:

- El progreso de A2 a B1 muestra un descenso de errores (de una media de 191,30 a 135,40) aunque esta disminución puede deberse a la evitación de estructuras.
- Los errores más abundantes afectan a la gramática (48,61%) y el léxico (29,37%) –frente al 14,19% de la pronunciación, y el 3,58% de la pragmática o el discurso)–; aproximadamente, un 4,45% son ambiguos, y el 49,21% se deberían a interferencia.

- Las incorrecciones del nivel léxico se concentran en los aspectos formales más que en los semánticos (que son más resistentes a la adquisición).
- Las errores gramaticales más frecuentes y generalizados afectan al artículo, la estructura de la oración, las concordancias y los tiempos de pasado; además, en B1 persisten los que se registran en pronombres, preposiciones y la subordinación.
- En la pronunciación, la interferencia suele persistir en B1 y quizá es donde más influye la L1, aunque algunos errores son generalizados, como la articulación de /r/.
- Los errores pragmático-discursivos revelan gran variación individual no siempre relacionable con el nivel, y quizá dependen de las propias habilidades del alumno; con todo, los chinos muestran relaciones de cohesión más pobres.

Entre las limitaciones de nuestro estudio, se pueden destacar las siguientes:

- Al basarse en datos orales, no es claro si ciertas desviaciones se deben a la competencia o la actuación, o respecto al tipo de dificultad o el nivel lingüístico.
- Solo reunimos datos de alumnos de nivel intermedio-bajo, y no es posible extraer conclusiones acerca de su capacidad para lograr una competencia casi bilingüe.
- No es posible generalizar los resultados dado el reducido número de alumnos participantes en cada grupo, aunque sí se vislumbran tendencias particulares.

Se pueden proyectar las siguientes tareas futuras de investigación:

- Ampliar el corpus con más participantes, más lenguas maternas y más niveles.
- Confirmar con pruebas complementarias (ej., escritas) el tipo o el nivel lingüístico de algunos fenómenos, o si se deben a la competencia o la actuación.
- Validar la taxonomía de errores entre dos anotadores independientes y con un subconjunto del cor-

pus, estimando el grado de coincidencia entre analistas y la consistencia del etiquetado (mediante el índice Kappa, κ; Herrera *et al.*, 2011: 194).

Si bien el alcance de este proyecto es reducido y los resultados no son todos generalizables, la principal aportación del estudio es un análisis de la producción y de los errores del español oral no nativo con una metodología rigurosa y extendible a otro corpus. Más aún, el banco de datos recogido ha sido procesado informáticamente y es accesible para la consulta por medio de una interfaz en línea (figuras 12 y 13). Creemos ciertamente que será aprovechable tanto por la comunidad investigadora en la adquisición de lenguas como por los docentes de español/LE.



Figura 12. Muestra de la pantalla inicial de la interfaz de consulta al corpus

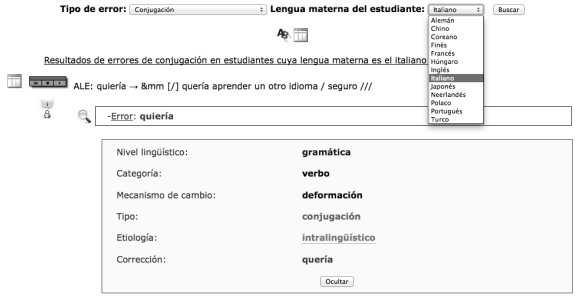


Figura 13. Muestra de consulta de errores mediante la interfaz informática en línea

BIBLIOGRAFÍA

BARLOW, M. (2005) «Computer-based analyses of learner language», en Ellis y Barkhuizen, *op. cit.*, chapter 14, pp. 335-357.

CAMPILLOS LLANOS, L. (2012) *La expresión oral en español lengua extranjera: interlengua y análisis de errores basado en corpus.* Tesis doctoral inédita. Universidad Autónoma de Madrid. Dpto. de Lingüística y Lenguas modernas.

CORDER, P. (1971/1991) «Idiosyncratic dialects and error analysis». *International Review of Applied Linguistics*, nº 9, 2, pp. 147-60. Traducción al español: «Dialectos idiosincrásicos y análisis de errores», en J. Muñoz Liceras (comp.) (1991) *La adquisición de las lenguas extranjeras*, pp. 63-77. Madrid: Visor.

CRESTI, E., y M. MONEGLIA (eds.) (2005). *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam: John Benjamins.

DAGNEAUX, E., S. Denness, y S. Granger (1998) «Computer-aided error analysis». *System*, 26(2), pp. 163-174.

DÍAZ-NEGRILLO, A., y J. FERNÁNDEZ DOMÍNGUEZ (2006) «Error tagging systems for learner corpora», *RESLA*, pp. 83-102. Disponible en: <http://dialnet.unirioja.es/ servlet/dcfichero_articulo?codigo=2198610&orden=72810> [11/01/2013]

DÍAZ-NEGRILLO, A., y S. VALERA, (2009) «What results give which strategies in error annotation?». *CALICO* Pre-Conference Workshop.
<www.sfs.uni-tuebingen.de/~dm/events/aall09/diaz-negrillo-valera.pdf> [11/01/2013]

DÖRNYEI, Z. (2007) *Research Methods in Applied Linguistics.* Oxford: Oxford University Press.

DULAY, H. C., y M. K. BURT (1976) «Creative construction in second language learning and teaching». *Language Learning,* número especial 4.

DUSSIAS, P. E. (2003) «Cognitive Perspectives on the Acquisition of Spanish as a Second Language». En B. Lafford y R. Salaberry (eds.) *Spanish Second Language Acquition,* capítulo 8, pp. 233-261.

ELLIS, R., y G. BARKHUIZEN (2005) *Analyzing Learner Language*. Oxford: Oxford University Press.

FERNÁNDEZ LÓPEZ, S. (1990) *Análisis de errores e interlengua en el aprendizaje del español como lengua extranjera.* Tesis doctoral. Universidad Complutense de Madrid. Facultad de Filología, Dpto. de Filología Románica.

GILQUIN, G. (2008) «Combining contrastive and interlanguage analysis to apprehend transfer: detection, explanation, evaluation». En Gaëtanelle Gilquin, Szilvia Papp y María Belén Díez-Bedmar (eds.) (2008) *Linking up contrastive and learner corpus research*. Amsterdam/New York, NY: Rodopi. Language and Computers - Studies in Practical Linguistics Series, 66.

GRANGER, S. (2003) «Error-tagged Learner Corpora and CALL: a promising synergy», *CALICO Journal*, 20 (3), pp. 465-480.

GRANGER, S. (2008) «Learner corpora». En Lüdeling & Kytö (eds.), *op. cit.*, pp. 259–275.

GRANGER, S., O. Kraifa, C. Pontona, G. Antoniadisa, y V. Zampa (2007) «Integrating learner corpora and natural language processing: A crucial step towards reconciling technological sophistication and pedagogical effectiveness». *ReCALL*, 19, pp. 252-268.

HEIFT, T. y M. SCHULZE (2007) *Errors and Intelligence in Computer-Assisted Language Learning: Parsers and Pedagogues.* London: Routledge.

HERRERA SOLER, H., R. Martínez Arias, y M. Amengual Pizarro (2011) *Estadística aplicada a la investigación lingüística.* Madrid: EOS Universitaria.

IZUMI, E., K. UCHIMOTO, y H. ISAHARA (2004) «SST Speech corpus of Japanese learners' English and automatic detection of learners' errors». *ICAME Journal,* nº. 28, pp. 31-48. Documento accesible en:
<http://icame.uib.no/ij28/Izumi.pdf> [11/01/2013]

IZUMI, E., K. Uchimoto y H. Isahara (2005) «Error Annotation for Corpus of Japanese Learner English». En *Proceedings of the Sixth International Workshop on Linguistically Interpreted Corpora (LINC 2005),* pp. 71-80. <http://aclweb.org/anthology-new/I/I05/I05-6009.pdf> [11/01/2013]

JAMES, C. (1998) *Errors in Language Learning and Use. Exploring Error Analysis.* London/New York: Longman. Applied Linguistics and Language Study Series.

KILGARRIFF, A., P. RYCHLY, P. SMRZ, y D. TUGWELL (2004) «The Sketch Engine». En *Proc. EURALEX 2004, Lorient, France,* pp 105-116. <http://www.sketchengine.co.uk>

KÖHLMYR, P. (2003) *«To Err is Human…» An investigation of Grammatical Errors in Swedish 16-year-old Learners' Written Production in English.* Göteborg: Acta Universitatis Gothoburgensis.

LARSEN-FREEMAN, D., y M. H. LONG (1991)[1994] *An Introduction to Second Language Acquisition Research.* London/New York: Longman. Applied Linguistics and Language Study.

LAFFORD, B., y M. R. SALABERRY (2003) *Spanish second language acquisition: state of the science.* Washington: Georgetown University Press.

LOTT, D. (1983) «Analysing and counteracting interference errors». *ELT Journal*, vol. 37/3, pp. 256-261.

LÜDELING, A., M. WALTER, E. KROYMANN, y P. ADOLPHS (2005) «Multi-level error annotation in learner corpora». *Proc. of the Corpus Linguistics Conference 2005.* <http://www2.huberlin.de/korpling/projekte/falko/FALKOCL2005.pdf?PHPSESSID=5bf51ebe3e6c9d694c16ffc9b27efc15> [11/01/2013]

LÜDELING, A., y M. KYTÖ (eds.) (2008) *Corpus Linguistics. An International Handbook*. 2 Vols. Berlin: Mouton de Gruyter.

MACKEY, A., y S. Gass (2005) *Second Language Research. Methodology and Design*. London/N.J.: Lawrence Erlbaum Associates.

MARCOS MARÍN, F. (1994) *Informática y humanidades.* Madrid: Gredos.

MCENERY, A., R. XIAO, y Y. TONO (2006) *Corpus-Based Language Studies. An Advanced Resource Book*. London: Routledge (Routledge Applied Linguistics).

MITCHELL, R., L. DOMÍGUEZ, Mª. ARCHE, F. MYLES y E. Marsden (2008) «Research Report: Linguistic Development in L2 Spanish: Creation and analysis of a learner corpus». En Roberts, L., F. Myles, y A. David. *EuroSLA Yearbook, 8.* Amsterdam; Philadelphia: John Benjamins, pp. 287-304. Documento disponible en: <http://www.splloc.soton.ac.uk/doc/SPLLOC_1_Report_Final.pdf> [11/01/2013]

MORENO SANDOVAL, A. (1998) *Lingüística computacional.* Madrid: Síntesis.

MORENO SANDOVAL, A., y GUIRAO, J. M. (2006) «Morpho-syntactic Tagging of the Spanish C-ORAL-ROM Corpus: Methodology, Tools and Evaluation». En Y. Kawaguchi, S. Zaima y T. Takagaki (eds.) *Spoken Language Corpus and Linguistic Informatics,* pp. 199-218. Amsterdam: John Benjamins. Una demostración de GRAMPAL está en: <http://cartago.lllf.uam.es/grampal/grampal.cgi> [11/01/2013]

MYLES, F. (2005) «Interlanguage corpora and second language research». *Second Language Research,* 21(4), pp. 373-391.

NEMSER, W. (1971) «Approximative systems of foreign language learners». *International Review of Applied Linguistics*, IX (2), pp. 115-123.

NICHOLLS, D. (2003) «The *Cambridge Learner Corpus* – error coding and analysis for Lexicography and ELT». En Archer *et al.* (eds.) *Proceedings of the Corpus Linguistics 2003 Confe-*

crence (CL2003). Lancaster University: University Centre for Computer Corpus Research on language, pp. 572-581. Disponible en: <http://ucrel.lancs.ac.uk/publications/CL2003/papers/nicholls.pdf> [11/01/2013]

NICKEL, G. (1981) «Aspects of error analysis: "errare humanum est"». *AILA Bulletin,* 29, pp. 1-27.

SELINKER, L. (1972) «Interlanguage», *International Review of Applied Linguistics*, vol. 10, nº 3, pp. 209□301. Traducción española: «La interlengua», en J. Muñoz Liceras (1992) *La Adquisición de las lenguas extranjeras: hacia un modelo de análisis de la interlengua.* Madrid: Visor. Pp. 79-101.

VALVERDE MATEOS, A. (2012) *Análisis de errores de aprendientes de Francés Lengua Extranjera (FLE) basado en corpus orales.* Tesis doctoral sin publicar. Madrid: Universidad Autónoma de Madrid. Dpto. de Lingüística y Lenguas modernas.

VÁZQUEZ, G. (1999) *¿Errores? ¡Sin falta!* Madrid: Edelsa.

PROGRAMAS

Transana 2.05. Universidad de Wisconsin. <www.transana.org> [11/01/2013]

NOTAS

[*] Trabajo financiado por la Comunidad de Madrid y el Fondo Social Europeo (contrato predoctoral).

[1] El acceso a la interfaz se encuentra disponible en: http://cartago.lllf.uam.es/corele/index.html

[2] En los errores interlingüísticos, marcamos información sobre de la(s) lengua(s) que quizá interfiere(n).

[3] Agradecemos al Dr. José Mª. Guirao de la Universidad de Granada el desarrollo y la puesta a disposición de dicha herramienta informática, que nos ha agilizado este proceso.

[4] En los cuantificadores indefinidos reunimos formas como *nadie, algo* o *un poco*.

[5] El recuento de auxiliares aquí presentado corresponde solamente a los que no pertenecen a las formas verbales complejas (tiempos compuestos, perífrasis o pasiva), las cuales se computan dentro de la categoría del verbo. Si contáramos estos casos, la producción total de auxiliares sería de 992.

[6] Los errores ambiguos tienen dos interpretaciones, y a veces dos diagnósticos distintos; p. ej., **creo [ˈkrei]* puede ser de conjugación (intralingüístico) o de pronunciación (por factores desconocidos).