CEDEL2: Corpus Escrito del Español como L2

CRISTÓBAL LOZANO
Universidad de Granada

RESUMEN

Debido al auge de estudios formales de adquisición del español L2 en los últimos años, CEDEL2 surge para dar respuesta a esta creciente necesidad investigadora y así proporcionar una gran cantidad de datos en español L2. CEDEL es parte del proyecto Word Order in Second Language Acquisition Corpora (WOSLAC), cuyo objetivo principal es determinar las propiedades que operan en las interfaces (léxico-sintaxis y sintaxis-discurso) y que restringen el orden de palabras en español L2.

ABSTRACT

Due to the increasing interest in L2 Spanish research, CEDEL2 is welcome as a new source of large-scale data for researchers of L2 Spanish. CEDEL2 is part of the project Word Order in Second Language Acquisition Corpora (WOSLAC) whose aim is to determine the properties at the interfaces (lexicon-syntax and synatx-discourse) that constrain word order in L2 Spanish.

PALABRAS CLAVE: *corpus de aprendices, adquisición del español, interfaces, learner corpora, Spanish second language acquisition,*

1 INTRODUCCIÓN

Con sus más de 20 millones de palabras, CHILDES (McWhinney 2000) se ha convertido en el corpus de referencia estándar más usado en los estudios del lenguaje infantil, especialmente de inglés L1. También han surgido recientemente extensos corpus nativos de cientos de millones de palabras en inglés, p.ej., *ICE*, *BNC*, *BYU-CAM* (véase Lüdeling et al. 2008, McEnery et al. 2005) y también en español, p. ej., *CREA*, *CORDE*, *Corpus del Español*, etc.

1.1 Corpus de aprendices de L2

La aparición del *International Corpus of Learner English* (ICLE) (Granger et al. 2002) a principios de los 90 marca el comienzo de los grandes corpus de aprendices. Los datos proceden de ensayos argumentativos escritos por alumnos universitarios aprendices de inglés L2 de diversas lenguas maternas. Para las comparaciones de interlengua vs. lengua nativa, se creó un corpus equivalente de inglés nativo *Louvain Corpus of Native English Essays* (LOCNESS).

Aparte de ICLE, existen otros dos grandes corpus de aprendices de inglés L2, aunque no están disponibles comercialmente: el *Longman Learner Corpus* (LLC) y el *Cambridge Learner Corpus* (CLC).

Siguiendo la estela de ICLE, se están creando dos corpus de aprendices (L1 español – L2 inglés): el *Written Corpus of Learner English* (WriCLE) de la Universidad Autónoma de Madrid y el *Santiago University Learner of English Corpus* (SULEC). Igualmente, ICLE ha dado pie a numerosos corpus de aprendices de inglés L2 con diversas L1 (véase Lüdelling et al. 2008). Existe también un corpus de L1 inglés – L2 francés en CHILDES, el *French Learner Language Oral Corpus* (FLLOC).

1.2 Corpus de aprendices de español L2

El creciente interés por el estudio de la adquisición del español L2 en los últimos años (p. ej., Lafford & Salaberry 2003, Montrul 2004) no ha venido acompañado por la creación de grandes corpus de español L2. En este contexto surge la creación de dos corpus de español L2: *Corpus Escrito del Español* (CEDEL2) en la Universidad Autónoma de Madrid y *Spanish Learner Language Oral Corpus* (SPLLOC) en la Universidad de Southampton (Mitchell et al. 2008).

2 CEDEL2 (CORPUS ESCRITO DEL ESPAÑOL L2)

CEDEL2 es un corpus escrito de L1 inglés – L2 español de todos los niveles (principiante, intermedio, avanzado), acompañado de un corpus nativo similar. CEDEL2 surge en el seno del proyecto *Word Order in Second Language Acquisition Corpora* (WOSLAC), dirigido por Amaya Mendikoetxea, Universidad Autónoma de Madrid (véase: http://www.uam.es/woslac y Chocano et al. 2007). El objetivo del programa de investigación viene marcado por recientes hallazgos en L2 (p. ej., Sorace 2005, 2006): determinar el papel que juegan las interfaces en el desarrollo de la interlengua (Lozano & Mendikoetxea 2007, en prensa).

3 METODOLOGÍA

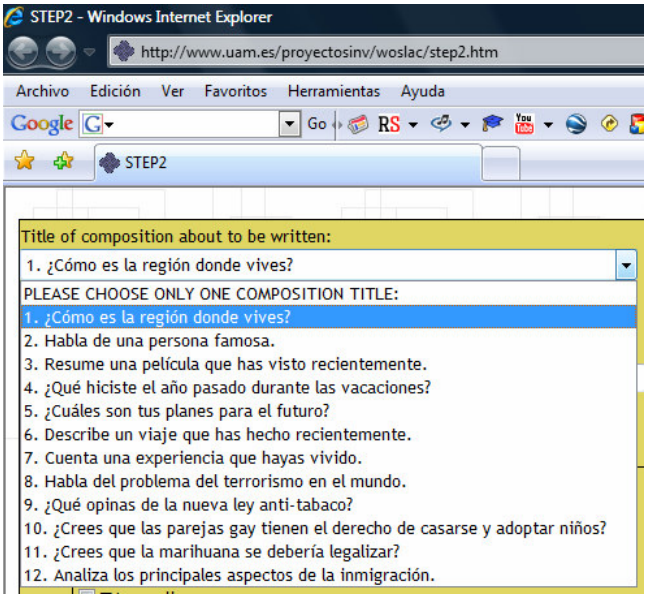
Los corpus de aprendices se suelen diseñar para que los aprendices produzcan determinadas estructuras lingüísticas en las que están interesados los investigadores. Sin embargo, en la creación de CEDEL2, optamos por ceñirnos a las recomendaciones de expertos en la creación de corpus, siguiendo los diez principios para la creación de corpus propuestos por Sinclair (2005).

3.1 Principios del diseño del corpus

Principio 1: Contenido del corpus. El contenido del corpus debe ser creado según criterios externos (la función comunicativa de los textos del corpus) y no criterios internos (los referidos a la lengua de los textos) (Sinclair 2005:1). A diferencia de otros corpus de aprendices diseñados para recoger determinadas estructuras lingüísticas, CEDEL2 sigue **criterios externos**: se diseñó el corpus para que no hubiese una desproporción de unas estructuras lingüísticas (o de léxico) sobre otras. Este principio está relacionado con el segundo.

Principio 2: *Representatividad.* El corpus debe ser lo más representativo posible de la lengua de la que ha sido escogido (Sinclair 2005: 2). Los aprendices de CEDEL2 pueden elegir libremente el tema de la redacción escrita de un total de **12 temas diferentes** (Ilustración 1) muestreados de manuales empleados en la enseñanza del español. Al representan varios temas (según grado de dificultad), se pretende elicitar todo tipo de estructuras lingüísticas y ninguna en particular.

Ilustración 1: Temas de redacción de CEDEL2



La representatividad también implica que debe haber muestras de la interlengua en diferentes estadios de desarrollo. En CEDEL2 han participado aprendices de **todos los niveles de competencia gramatical** (principiante, intermedio, avanzado) según un test de nivel estandarizado (Universidad de Winconsin 1998) que se habilitó para que los aprendices pudiesen completarlo directamente *online* (Apéndice 2). Finalmente, la representatividad está relacionada con el muestreo del corpus: *longitudinal* o *transversal* ('cross sectional'). Al ser logísticamente difícil tomar un diseño longitudinal, en CEDEL2 se optó por un **diseño transversal** estándar, basado en distintos niveles de competencia.

Principio 3: Contraste. Sólo deben ser contrastados aquellos componentes del corpus que han sido diseñados para ser contrastados independientemente (Sinclair 2005:3). En CEDEL2 se están recogiendo datos para la creación de un **subcorpus de hablantes nativos** de español (véase la sección 4). Para que el contraste entre el subcorpus de aprendices y el de nativos sea lo más fiable posible, los nativos siguen las mismas pautas que los aprendices. Adicionalmente, CEDEL2 permite **contrastar dos interlenguas** (por ejemplo, niveles intermedio vs. avanzado).

Principio 4: Criterios estructurales. Los criterios para determinar la estructura de un corpus deben ser reducidos en número y claramente separables los unos de los otros (Sinclair 2005:5). Este principio es de suma importancia en la creación de extensos corpus nativos del tipo *monitor corpora*. Dado que CEDEL2 es un corpus de aprendices, nuestro criterio estructural más importante es la división del corpus completo en un **subcorpus de aprendices** (dividido en tres niveles) y un **subcorpus de nativos**, como queda apuntado.

Principio 5: Etiquetado. Cualquier información acerca del texto (aparte de la información alfanumérica: palabras y signos de puntuación) debería ser almacenada separadamente del texto puro para posteriormente ser fusionada con el texto si la aplicación informática lo requiere (Sinclair 2005:5). Como se detallará en la sección

3.3, nuestro etiquetador (*UAM CorpusTool*) deja el fichero de texto puro intacto y crea un nuevo fichero XML que contiene las etiquetas.

Principio 6: Muestra. Las muestras de la lengua del corpus al ser posible deberían consistir en documentos o transcripciones de eventos del habla completos. Esto implica que las muestras diferirán en tamaño sustancialmente (Sinclair 2005:7). Conviene destacar que:

> 'There is no virtue from a linguistic point of view in selecting samples all of the same size. True, this was the convention in some of the early corpora, and it has been perpetuated in later corpora [...] it is difficult to justify the continuation of the practice. The integrity and representativeness of complete artifacts is far more important than the difficulty of reconciling texts of different dimensions.' (Sinclair 2005:6).

En este sentido, en CEDEL2 encontramos **sólo textos completos**, independientemente del número de palabras que contengan.

Principio 7: Documentación. El diseño y la composición de un corpus debería ser documentado detalladamente con información sobre los contenidos y los argumentos que justifican las decisiones tomadas para que, si en un momento dado el investigador obtiene resultados 'extraños' y contra-intuitivos, se pueda comprobar si existe un error de diseño o de selección de textos (Sinclair 2005:8). En este sentido, en CEDEL2 se ha recogido **información detallada sobre diferentes aspectos de cada aprendiz** (véase la sección 3.2 Recogida de datos).

Principio 8: Equilibrio. El diseñador de corpus debe de tener como nociones meta la representatividad y el balance. Aunque estos objetivos no sean precisamente definibles y alcanzables, deben servir de guía en el diseño del corpus (Sinclair 2005:9). Sinclair se refiere a que el corpus debe estar equilibrado y contener muestras representativas de todo tipo de lengua (oral y escrito). De nuevo, este principio es relevante para los llamados *monitor corpora*. CEDEL2 es un **corpus escrito** y, como tal, los resultados sólo se extrapolarán a la interlengua escrita.

Principio 9: Tema. Cualquier control en el tema del corpus debería ser regido por criterios externos y no criterios internos (Sinclair 2005: 10). Este principio ya ha sido anteriormente: principio 1 (contenido) y principio 2 (representatividad). Como queda dicho, los temas de redacción están sopesados para dar lugar a un lenguaje lo más representativo posible.

Principio 10: Homogeneidad. El objetivo del corpus es alcanzar la homogeneidad de sus componentes y, al mismo tiempo, mantener una cobertura adecuada y evitar los textos atípicos (*rogue texts*) (Sinclair 2005:14). Una vez finalizada la recogida de los datos de CEDEL2 y previo al etiquetado de datos, los investigadores examinarán cada texto para poder descartar los textos que no se adecuen a los criterios del corpus (*rogue texts*).

3.2 Recogida de datos

Los datos de CEDEL2 están siendo recogidos *online* a través de formularios electrónicos (http://www.uam.es/woslac/start.htm): (1) *learning background*, (2) *placement test* y (3) *composition*. Según el principio 7 (documentación) visto arriba, se documentó la siguiente información:

(i) *Formulario 'learning background'* (Apéndice 1), donde se detalla:
   a. Datos personales (edad, sexo, estudios).
   b. Datos lingüísticos (lengua materna, lengua de los padres, lengua hablada en casa, edad de inmersión en español, estancias en países de habla hispana).
   c. Autovaloración de nivel estimado en español L2.
   d. Autovaloración de nivel estimado en otras L2.

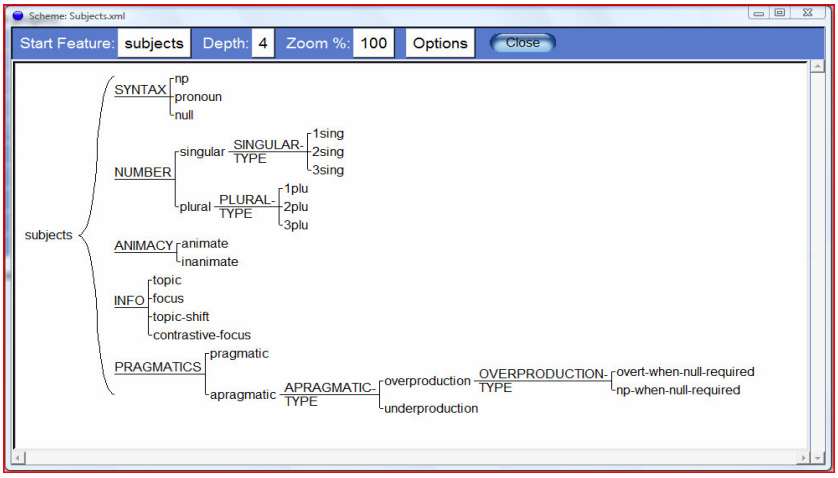
(ii) *Formulario 'composition'* (ver el Apéndice 3), donde se proporciona la redacción en sí y, además la siguiente información:
   a. Lugar de redacción (en clase o fuera de clase).
   b. Investigación previa sobre el tema y herramientas empleadas (intenet, periódicos, televisión, etc).
   c. Herramientas lingüísticas empleadas (diccionarios bilingües, monolingües, correctores ortográficos, ayuda nativa, etc).

Al hacer un análisis cuantitativo y encontrar resultados 'extraños, contra-intuitivos y conflictivos' (según el principio 10 de Sinclair), los datos cualitativos proporcionados en (i) y (ii) permiten al investigador indagar sobre las posibles causas de de tales resultados.

3.3 Etiquetado de datos

El software empleado para etiquetar CEDEL2 es el software gratuito *UAM CorpusToo* (O'Donnell 2007 y http://www.wagsoft.com/CorpusTool). Esta herramienta permite al lingüista seleccionar segmentos de un texto y anotarlo (etiquetarlo) según un esquema creado previamente por el investigador de acuerdo a sus necesidades. El texto etiquetado se guarda aparte en formato XML. CEDEL2 se ha empezado a etiquetar en relación a los pronombres sujeto (pronombres nulos y plenos) según un esquema (Ilustración 2): cada segmento (es decir, cada pronombre sujeto en este caso) es etiquetado según su sintaxis, número, animacidad, etc.

Ilustración 2: CEDEL2: Esquema de etiquetado de los pronombres sujeto



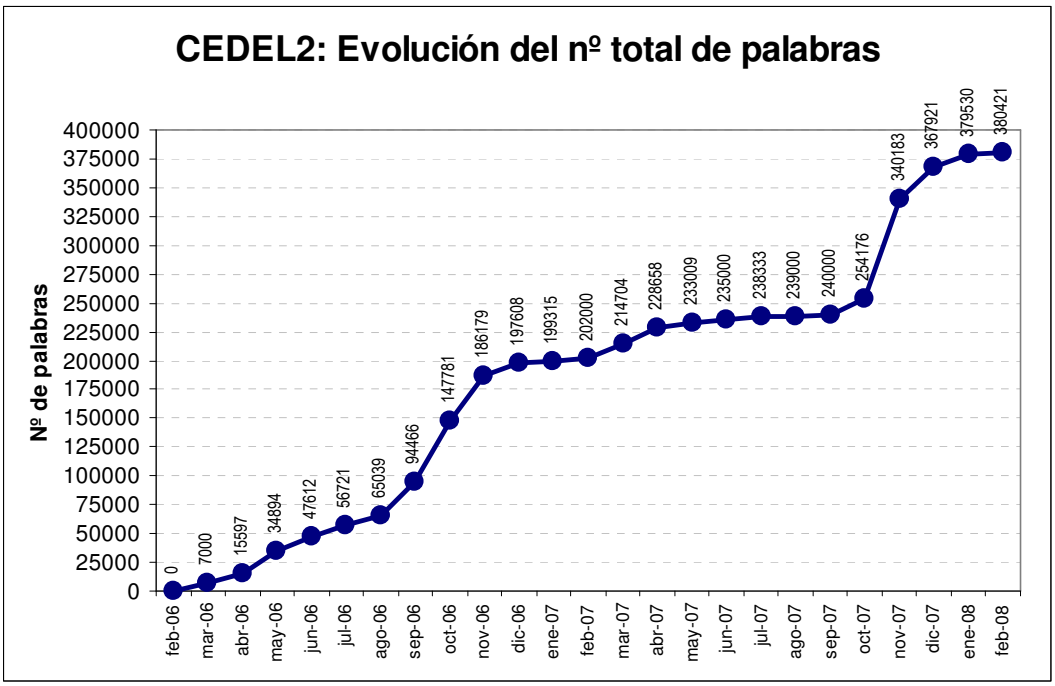
4 DATOS ACTUALES DE CEDEL2

En esta sección presentamos la distribución de los datos recogidos hasta la fecha (sección 4.1) y su procedencia (sección 4.2).

4.1 Distribución de los datos

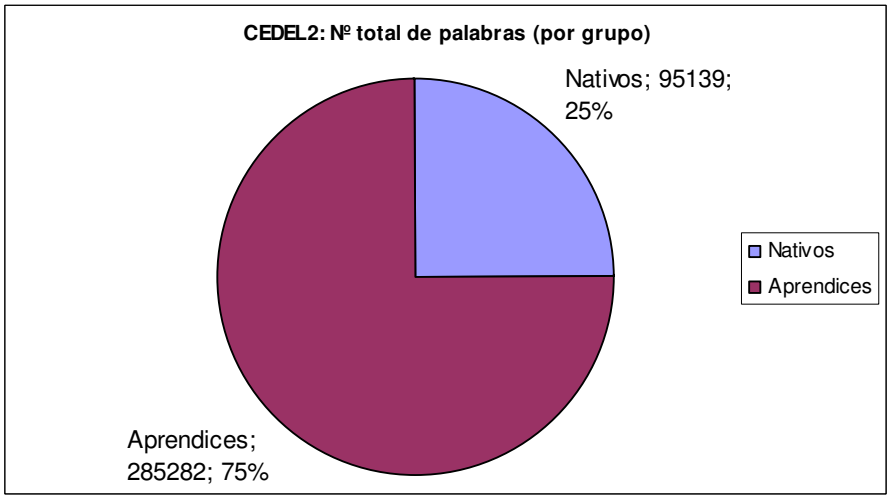
CEDEL2 consta de casi 400.000 palabras (febrero de 2008), Ilustración 3. Se puede apreciar en la tendencia creciente de la serie temporal dos fuertes subidas ocasionadas por los anuncios de CEDEL2 publicados en listas de distribución como *Linguist List*, *Infoling* y *AESLA*. El objetivo del proyecto de investigación WOSLAC es alcanzar el millón de palabras en CEDEL2.

Ilustración 3: CEDEL2: Total de palabras



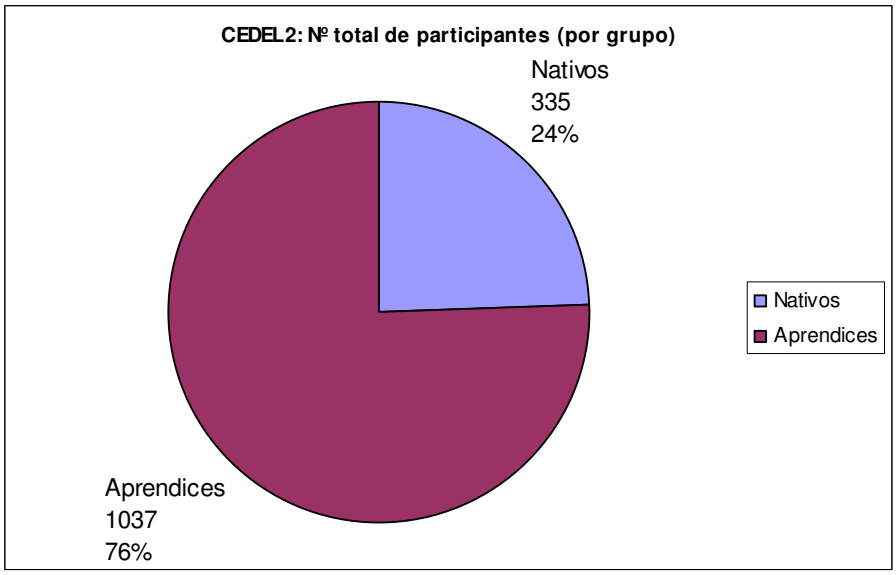
Un cuarto del número total de palabras del corpus proviene del subcorpus nativo (95.139 palabras, el equivalente a 25%), mientras que el resto procede del subcorpus de aprendices (285.282 palabras, 75%), Ilustración 4. Asumiendo que esta proporción se mantenga, al alcanzar el millón de palabras el subcorpus nativo alcanzaría las 250.000 palabras (similar a las 235.000 de LOCNESS) y el de aprendices alcanzaría 750.000 palabras (superior a las 200.000 palabras del subcorpus de español del ICLE).

Ilustración 4: Nº total de palabras (y porcentaje equivalente)



Nótese que los porcentajes anteriores son similares al porcentaje del número total de participantes de cada subcorpus, lo que implica un equilibrio en cada corpus entre la proporción de participantes y la proporción del número de palabras producidas (Ilustración 5).

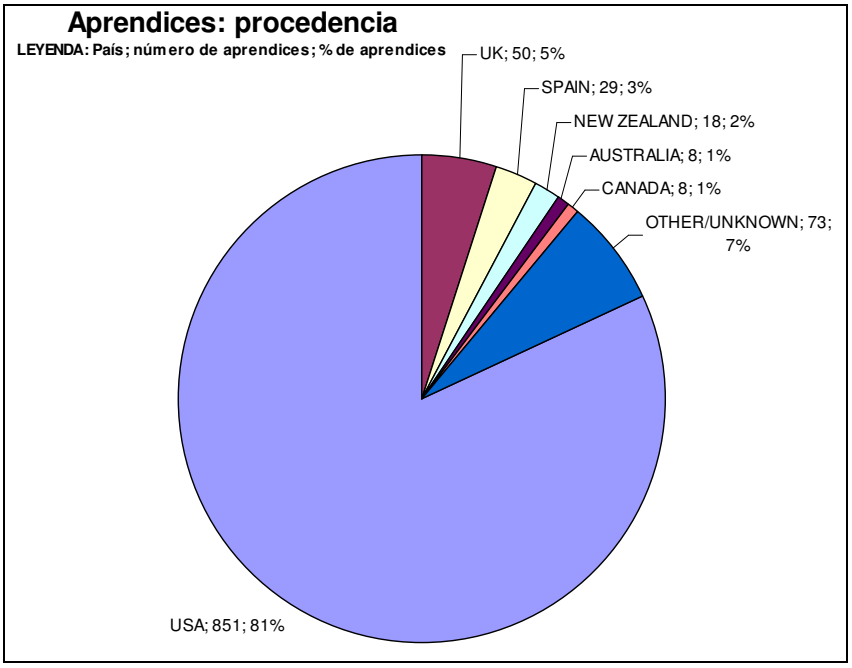
Ilustración 5: Nº total de participantes (y porcentaje equivalente)



4.2 Procedencia de los datos

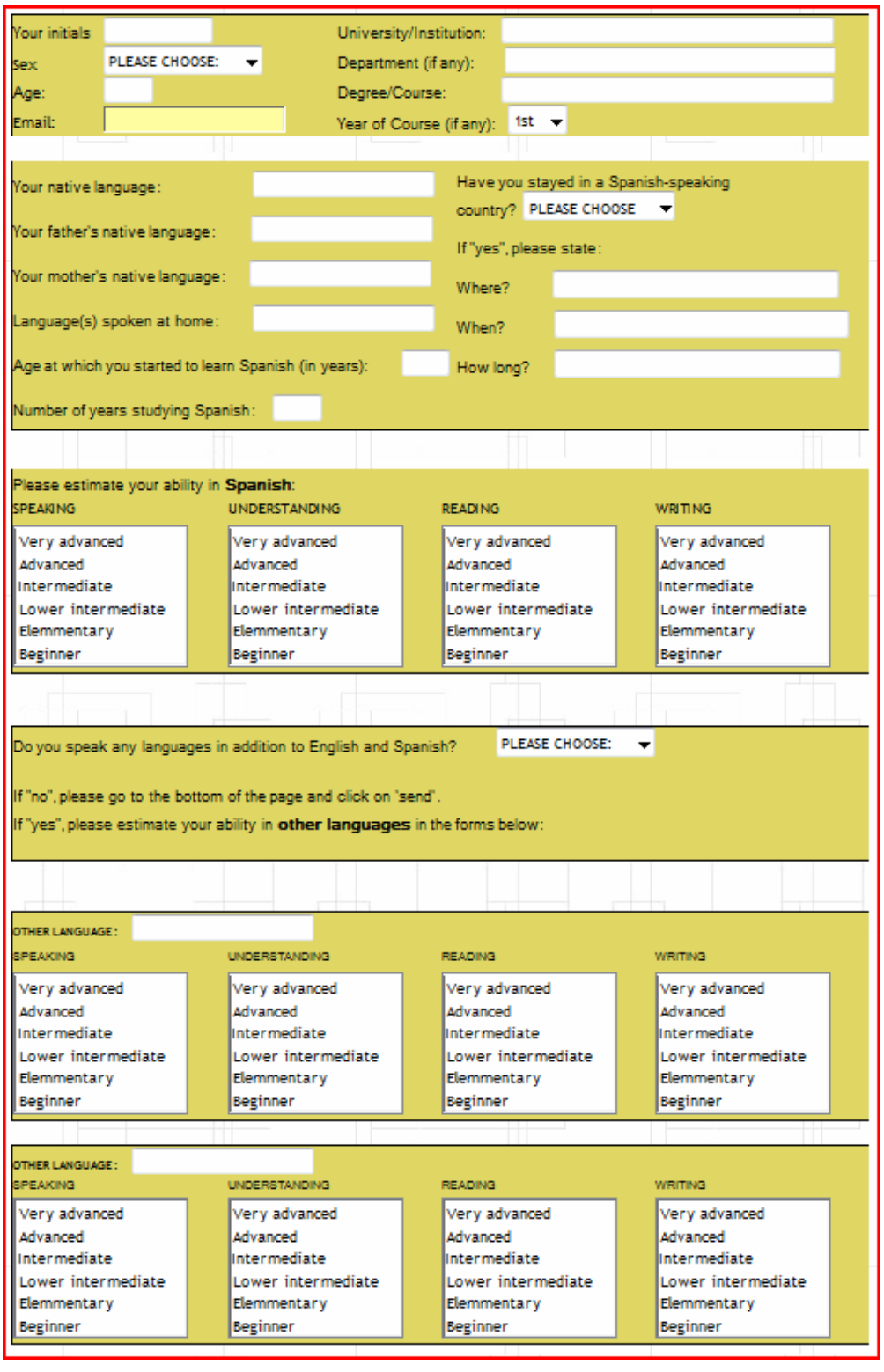
La gran mayoría de los datos de los aprendices provienen de estudiantes de español en diferentes universidades y colegios de EEUU, Ilustración 6, con un total de 851 participantes que representan el 81% del total.

Ilustración 6: Procedencia de los datos de CEDEL

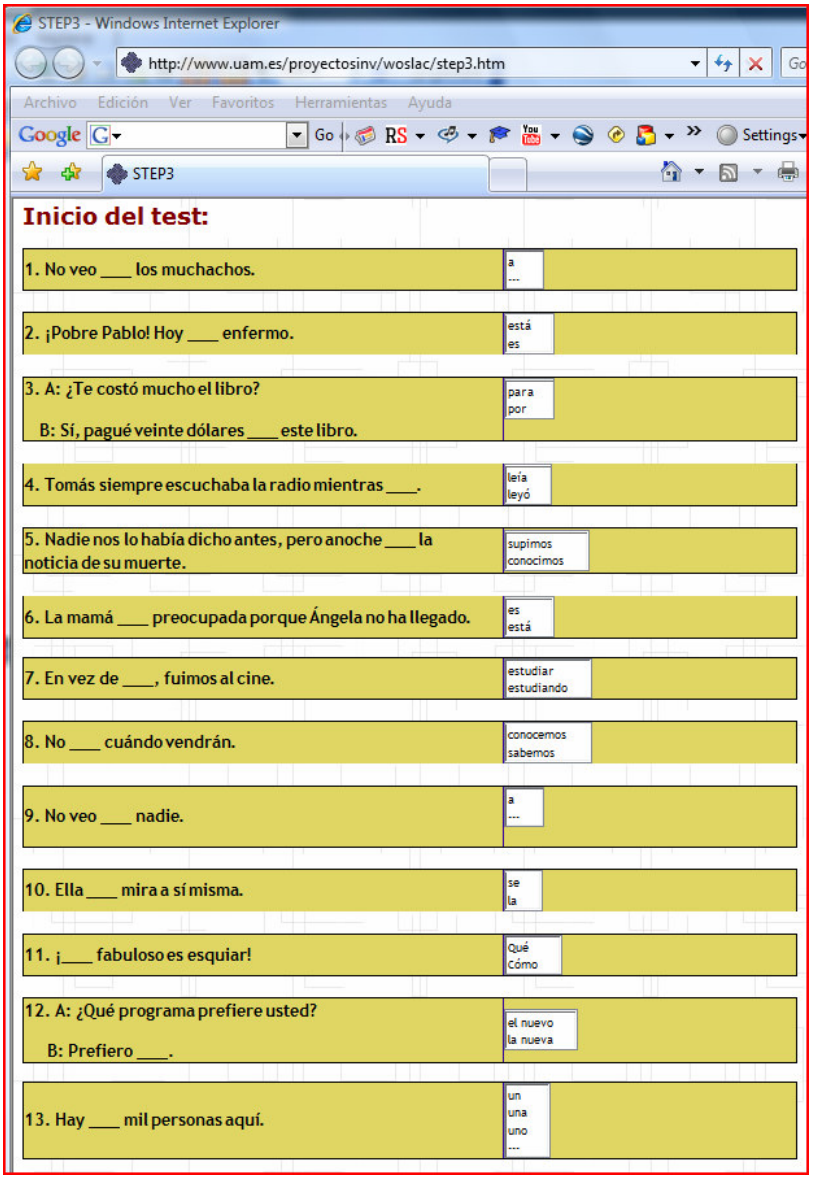


5 CONCLUSIÓN

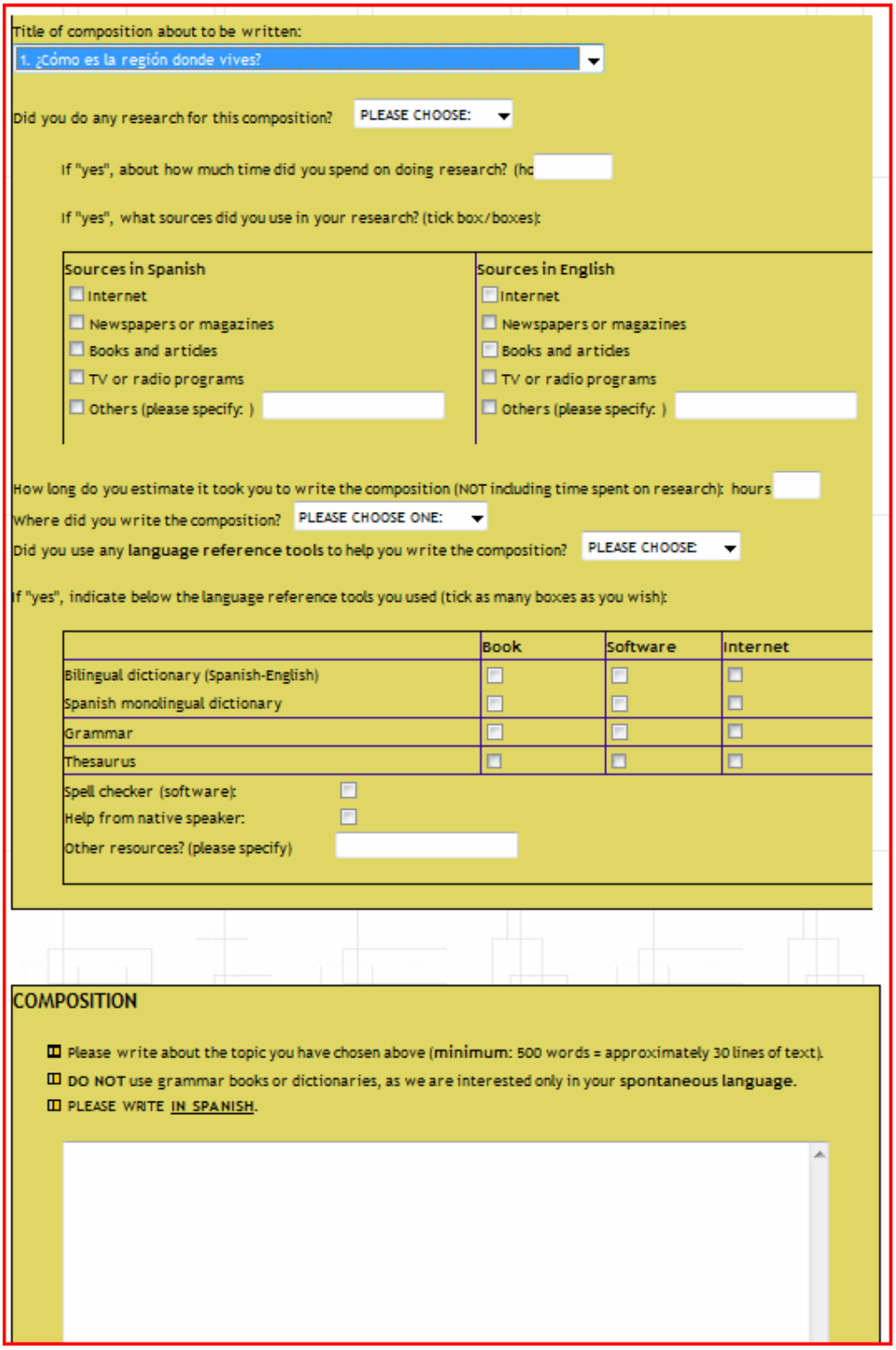
CEDEL2 (Corpus Escrito del Español L2) es un corpus que se está confeccionando para dar respuesta a las crecientes necesidades de investigación en el ámbito de español L2. El corpus consta de cerca de 400.000 palabras y se espera alcance un millón al final del periodo de investigación. CEDEL2 estará disponible gratuitamente en Internet para ser utilizado como fuente de datos por los investigadores de español L2 y como fuente de ejemplos para profesionales de ELE y aprendices del español.



Apéndice 1: Formulario online nº1 'learning background'



Apéndice 2: Formulario online nº 2 'placement test' (sólo hasta la pregunta nº 13)



Apéndice 3: Formulario online nº 3 'composition'

REFERENCIAS

Corpus Escrito Del Español L2 (CEDEL2) < http://www.uam.es/woslac/cedel2.htm > [Fecha de la consulta: 19-02-2008]

Chocano, G., Jiménez, R., Lozano, C., Mendikoetxea, A., Murcia, S., O'Donnell, M., Rollinson, P. y Teomiro, I. 2007. "An exploration into word order in learner corpora: The WOSLAC Project". Eds. M. Davies, P. Rayson, S. Hunston, y P. Danielsson. *Proceedings of the Corpus Linguistics Conference 2007*. Birmingham: University of Birmingham.

French Learner Language Oral Corpus (FLLOC). < http://www.flloc.soton.ac.uk > [Fecha de la consulta: 19-02-2008]

Granger, S., Dagneaux, E., and Meunier, F. 2002. *International Corpus of Learner English*. Louvain: UCL Presses Universitaires de Louvain.

Louvain Corpus Of Native English Essays (LOCNESS). < http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/locness1.htm > [Fecha de la consulta: 19-02-2008]

Lozano, C. en prensa. "Selective deficits at the syntax-discourse interface: Evidence from the CEDEL2 corpus". Eds. Y.-I. Leung, N. Snape y M. Sharwood-Smith. *Representational Deficits in SLA*. Amsterdam: John Benjamins.

Lozano, C. y Mendikoetxea, A. en prensa. "Postverbal subjects at the interfaces in Spanish and Italian learners of L2 English: a corpus análisis". Eds. G. Gilquin, S. Papp y B. Díez. *Linking up contrastive and corpus learner research*. Amsterdam: Rodopi.

Lozano, C. y Mendikoetxea, A. 2007. "Learner corpora and the acquisition of word order: A study of the production of Verb-Subject structures in L2 English". Eds. M. Davies, P. Rayson, S. Hunston, y P. Danielsson. *Proceedings of the Corpus Linguistics Conference 2007*. Birmingham: University of Birmingham.

Lüdeling, A., Kytoe, M., and McEnery, T. eds. 2008. *Corpus Linguistics: An International Handbook*. Berlin: Mouton de Gruyter.

McEnery, T. Xiao, R., and Tono, Y. 2005. *Corpus-based Language Studies: An Advanced Resource Book*. London: Routledge.

Mitchell, R., Domínguez, L., Arche, M., Myles, F., Marsden, E., enviado 2008. "SPLLOC: A new corpus for Spanish second language acquisition research". *EUROSLA Yearbook 8*.

Sinclair, J. 2005. "Corpus and text – Basic principles". Ed. M. Wynne. *Developoing Linguistic Copora: A guide to good practice.* Oxford: Oxbow Books.

Sorace, A. 2006. "Possible manifestations of shallow processing in advanced second language speakers". *Applied Psycholinguistics* 27, 88-91.

Sorace, A. 2005. "Selective optionality in language development". Eds. L. Cornips y K. P. Corrigan. *Syntax and variation: Reconciling the biological and the social*. Amsterdam: John Benjamins. 55-80.

Spanish Learner Language Oral Corpus (SPLLOC) [en línea] < http://www.splloc.soton.ac.uk > [Fecha de consulta: 19-02-2008]

University of Wisconsin, 1998. *The University of Wisconsin College-Level Placement Test: Spanish (Grammar) [Form 96M]*. Madison, WI: University of Wisconsin Press.

Word Order in Second Language Acquisition Corpora (WOSLAC), Universidad Autónoma de Madrid [en línea] < http://www.uam.es/woslac > [Fecha de consulta: 19-02-2008]