Procesamiento del lenguaje Natural, nº 41 (2008), pp. 297-298 recibido 7-05-2008; aceptado 16-06-2008

CHIEDE
Corpus de Habla Infantil Espontánea del Español

CHIEDE
Spontaneous Child Language Corpus of Spanish

Marta Garrote Salazar
Laboratorio de Lingüística Informática
Universidad Autónoma de Madrid
Campus de Cantoblanco,
Ctra. de Colmenar Viejo, Km. 15
28049-Madrid
marta.garrote@uam.es

José María Guirao Miras
ETSIIT; Dpto. de Lenguajes y Sistemas Informáticos
Universidad de Granada
C/ Periodista Daniel Saucedo Aranda s/n Granada
jmguirao@ugr.es

Resumen: El presente trabajo consiste en la demostración del funcionamiento de la página web desarrollada para la presentación y difusión del corpus de habla infantil CHIEDE.
Palabras clave: Corpus, lengua oral espontánea, lenguaje infantil, página web.

Abstract: This work consists on the demonstration of the web site developed for the presentation and spreading of the child language corpus CHIEDE.
Keywords: Corpus, spontaneous oral language, child language, web site.

El Corpus de Habla Infantil Espontánea del Español, CHIEDE, consta de unas 60.000 palabras. Aproximadamente un tercio del corpus está formado por habla infantil y los dos tercios restantes por habla adulta. La principal característica de CHIEDE es la espontaneidad de las interacciones que lo integran: las grabaciones se han llevado a cabo en su contexto natural. El recurso se presenta en diferentes formatos: una transcripción ortográfica, una transcripción fonológica automática, una versión etiquetada en XML y el alineamiento del texto y el sonido. Además, aportamos los resultados obtenidos tras la extracción, mediante métodos estadísticos, de la información de los textos anotados. El diseño del corpus presenta el siguiente aspecto:

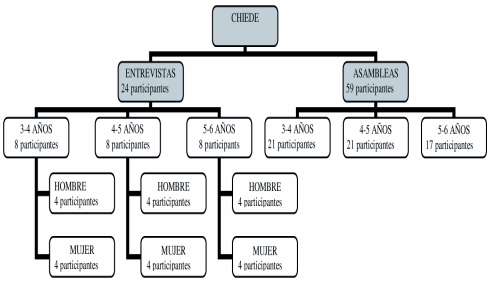

Figura 1: Diseño del corpus

CHIEDE cumple con todas las características que debe poseer un corpus de lengua oral actual. Su formato es electrónico, permitiendo el almacenamiento y la manipulación de los datos y su posible intercambio con otros investigadores interesados. Por su diseño proporcionado y su diversidad —variables de sexo, edad y situación comunicativa— garantiza una representatividad de la variedad lingüística en cuestión. Posee una estructura interna de clasificación de datos que posibilita una óptima explotación de los mismos Su

ISSN: 1135-5948

presentación en una página web (http://drusila.lllf.uam.es/chiede)[1] facilita su disponibilidad para todo aquel que esté interesado en su consulta. De esta forma, se puede acceder tanto a las transcripciones (alineadas con su sonido) como a cualquier tipo de información extraída de las mismas (número de palabras, listados de frecuencias). La página web de CHIEDE consta de los siguientes apartados:

- **Inicio**: <http://drusila.lllf.uam.es/chiede>.
- **Introducción**: en la que se explican los motivos que nos llevaron a realizar el presente proyecto y se plantea el estado de la cuestión sobre la lingüística de corpus y la ontogénesis del lenguaje.
- **Diseño del corpus**: en este apartado se describen el diseño del corpus y sus características.
- **Objetivos**: donde se detallan los objetivos de nuestro proyecto.
- **Transcripciones**: en esta sección facilitamos el texto de las transcripciones (tanto en su versión ortográfica como fonológica) alineado con su correspondiente sonido. Esto permite que en caso de duda respecto de las convenciones de transcripción, el usuario pueda escuchar la grabación original y juzgar por sí mismo.
- **Resultados**: todos los datos obtenidos de forma automática con métodos estadísticos pueden ser consultados aquí. Se facilitan pues listados de frecuencias por lemas, categorías, longitud media de enunciado (LME), etc.
- **Guías**: se facilitan al usuario dos guías básicas para la comprensión y utilización del corpus. En la primera de ellas se incluyen todas las convenciones de transcripción para la comprensión de los textos transcritos; la segunda contiene el *tagset* o sistema de etiquetado categorial con especificaciones sobre los criterios seguidos para el establecimiento del mismo.
- **Consulta**: la aplicación *Concordancias* permite al usuario consultar cualquier palabra. De esta forma, se pueden obtener todos los ejemplos de dicha palabra que aparezcan en el corpus, junto con su contexto y sonido.

Hasta la fecha, el corpus CHIEDE se ha utilizado como fuente de estudio de varias investigaciones sobre lenguaje infantil. Es nuestra intención en un futuro ampliar el tamaño del corpus, el número de participantes y la variedad de situaciones comunicativas.

Bibliografía

González Ledesma, A. y Garrote Salazar, M. 2007. Los marcadores discursivos en CHIEDE, un corpus de habla infantil espontánea. *Actas del XXII Congreso Internacional de la Asociación de Jóvenes Lingüistas*.

Garrote, M. y Moreno Sandoval, A. 2008. CHIEDE, a spontaneous child language corpus of Spanish. *Proceedings of the 3rd International LABLITA Workshop in Corpus Linguistics.*

Garrote, M., Guirao, J.M. y Moreno Sandoval, A. 2008. Extracción de variants léxicas en adultos y niños de un corpus de lengua oral espontánea. *Actas del 8º Congreso de Lingüística General.*

[1] Esta investigación ha sido parcialmente financiada por el proyecto Búsqueda de Respuestas Avanzada Multimodal y Multilingüe: Recursos Lingüísticos, (CICYT-TIN2007-67407-C03-02).