

LA METODOLOGÍA *RETAMHE* Y EL PROYECTO *CHILDES*: BREVIARIO PARA LA CODIFICACIÓN Y ANÁLISIS DEL LENGUAJE INFANTIL

Eliseo Diez-Itza, Catherine E. Snow* y Brian MacWhinney**
Universidad de Oviedo, * Universidad de Harvard y ** Universidad Carnegie

En este artículo se investigan algunas cuestiones metodológicas relativas al estudio del lenguaje infantil y se presenta la instrumentación y el software desarrollados en el proyecto *CHILDES* (Sistema de Intercambio de Datos del Lenguaje Infantil). El sistema proporciona potentes herramientas para investigar en el marco de la metodología de Registro, Transcripción y Análisis de Muestras de Habla Espontánea (*RETAMHE*). Se proporciona información del sistema de codificación (*CHAT*) que permite introducir finas distinciones en el proceso de transcripción, al tiempo que se exponen los requerimientos básicos (*minCHAT*) para crear archivos *CHAT*. El paquete de programas (*CLAN*) específicamente diseñados para analizar esos archivos que contienen transcripciones de muestras de habla, incluye recuentos de frecuencias, búsqueda de palabras, análisis de la interacción, etc. El núcleo de comandos y opciones de *CLAN* constituye el sistema *minCLAN*, del que se ofrecen algunas pautas. Se trata con todo ello de facilitar los primeros pasos para aprender a usar los instrumentos metodológicos de *CHILDES*.

RETAMHE methodology and the CHILDES project with a focus on minimal standards to codify and analyze child language. This paper investigates some methodological issues concerning research in the field of child language and introduces a set of instruments and software developed in the context of the *CHILDES* (Child Language Data Exchange System) Project. The *RETAMHE* system provides powerful tools for recording, transcribing, and analyzing spontaneous speech samples. Information about the use of the coding system (*CHAT*) for making distinctions in the transcription process is provided together with the minimum set of standards (*minCHAT*) to create *CHAT* files. The *CLAN* computer programs are used to analyze files containing transcriptions of speech samples for frequency counts, word searches, interactional analyses, etc. The core set of *CLAN* commands and options constitute the *minCLAN* system. The first steps in learning to use the *CHILDES* tools are also explained.

El devenir histórico de la investigación científica del lenguaje infantil ha venido

Correspondencia: Eliseo Diez-Itza
Facultad de Psicología
Universidad de Oviedo
33003 Oviedo (Spain)
E-mail: ditza@correo.uniovi.es

marcado, como no podía ser de otra manera, por inflexiones en la metodología. Así, desde sus orígenes más recientes, ligados a los de la psicología del desarrollo, se podrían distinguir tres etapas, presididas por otras tantas orientaciones metodológicas. En la primera de ellas, los naturales afanes des-

criptivos cobraron forma en la metodología observacional de los diarios. La segunda tradujo el interés diferencial en el método correlacional. Y la tercera, en la que ya estaba en juego la explicación del lenguaje infantil, vio proliferar el método experimental.

El estudio moderno del lenguaje infantil, que cumple ahora un cuarto de siglo si tomamos la referencia obligada de quien fue su destacado precursor (Brown, 1973), tampoco puede entenderse al margen de cambios metodológicos de naturaleza fundamentalmente tecnológica. En torno a esa fecha se produce una eclosión de investigaciones del lenguaje infantil marcadas por un nuevo enfoque: la idea de que el lenguaje y su desarrollo no pueden ser aislados de su contexto conversacional, es lo que en otro lugar hemos llamado el «giro comunicativo» (Diez-Itza, 1992).

El giro comunicativo, que llegó de la mano de los enfoques funcionales, socioculturales y pragmáticos, vino a decir que estábamos haciendo las fotos con las figuras cortadas por la mitad, que no podíamos hablar del lenguaje del niño, de un niño, que el lenguaje es cosa de dos, que el discurso es compartido y que, de algún modo, había que volver atrás, volver a observar y a describir la adquisición y desarrollo del lenguaje en sus escenarios comunicativos naturales. Muy gráficamente lo expresó Bruner, uno de sus principales protagonistas, desdiciendo por artificial el elegante laboratorio de Oxford «*a favor del desorden de la vida del hogar, fuimos hacia los niños en lugar de hacerlos venir hacia nosotros*» (Bruner, 1986, p. 13). Viejos métodos (la observación en el hogar), para nuevos enfoques (las variables pragmáticas: las condiciones de uso).

Se volvió pues al inevitable e insustituible método observacional, pero en unas condiciones tecnológicas impensables para nuestros tatarabuelos en estas lides, que seguían a sus hijos lápiz en ristre. El grafito se convirtió en mágica banda magnética y el

resto resultaba bastante simple: «*Cada quincena, nosotros, uno o dos observadores y yo, o los otros dos juntos, visitábamos sus hogares durante cerca de una hora y grabábamos media hora de cintas de video y de audio de la madre y el niño jugando a lo que en ese momento resultaba ser su actividad recreativa. [...] Las cintas de video y de audio eran transcritas tan pronto como era posible después de grabarlas, habitualmente en una semana, ayudados por notas y por la memoria. Las transcripciones tenían la forma habitual de tres columnas de anotaciones ordenadas temporalmente, en las que se daban los enunciados maternos, los enunciados del niño y las descripciones del contexto*» (Bruner, 1986, p. 48).

Esta es en términos sencillos la metodología de «Registro, Transcripción y Análisis de Muestras de Habla Espontánea», que abreviadamente denominamos metodología RETAMHE (Diez-Itza, 1992) y que explicamos con mayor pormenor del habitual en el libro que se cita, ya que suele darse por sentado el procedimiento en estos casos. De todos modos, el ámbito español constituye en tal sentido una excepción merced sobre todo a la obra de Miguel Siguán y sus colaboradores (Arnau y Boada, 1985; Serra, 1976; Siguán, 1983; Siguán, Colomina y Vila, 1990; Solé, 1988; Triadó y Forns, 1989). Dicha metodología se convirtió en el paradigma moderno de la investigación del lenguaje infantil, desde que Brown (1973) y sus colaboradores de Harvard empezaran el registro longitudinal de un niño de poco más de dos años (Adán), el ocho de octubre de 1962, al que siguieron casi paralelamente los de otras dos niñas (Eva y Sara) y las consiguientes transcripciones. El deseo que Brown expresa al final del prefacio de su obra se ha visto sobradamente cumplido y en gran parte se debe a su iniciativa: «*habiendo sido psicólogo durante veinte años he vivido entre unas cuantas prometedoras y activas tradiciones investigadoras. Por*

desgracia muchas de ellas no han dejado huellas importantes en forma de nuevo conocimiento. El desarrollo del lenguaje es hoy un campo de investigación prometedor y activo. Mi más profundo deseo es que nos deje un claro incremento del conocimiento psicológico» (Brown, 1973, p. 20).

Si la metodología RETAMHE ha contribuido decisivamente al estudio y conocimiento de la adquisición y desarrollo del lenguaje, no se debe sólo al trabajo pionero de Brown, sino también a su generosidad. Antes que sus resultados, compartió sus datos con todo aquel que se lo pidió y de ellos surgieron infinidad de estudios y todavía hoy día el «corpus» de Brown sigue siendo objeto de investigación, como más adelante se pondrá de manifiesto. La ingente tarea que supone la recogida de datos de acuerdo con esta metodología alcanza una rentabilidad científica inusitada cuando los datos obtenidos se comparten. Si el «giro comunicativo» se caracterizó por entender el lenguaje como una actividad compartida, su actitud metodológica no le fue a la zaga convirtiendo la investigación en una tarea compartida por la comunidad científica internacional en sentido estricto. Al complejo trabajo de transcripción, Brown añadió el traslado de las extensísimas transcripciones de Adán, Eva y Sara a un cliché que permitió tirar un número limitado de copias que se distribuyeron por todo el mundo, hasta el punto de que Brown al final sólo conservaba la copia original archivada ya como un documento histórico.

El progreso tecnológico había permitido desde los años 60 grabar conversaciones infantiles y reproducir, aunque fuera de un modo restringido, sus transcripciones, pero la idea de compartir tales datos del lenguaje infantil, no encontró su vía ilimitada en posibilidades hasta el desarrollo de la informática y sus ventajas de copia instantánea e incluso, como luego ha ocurrido, de envío a distancia. Con el propósito de aprovechar

estos nuevos recursos tecnológicos creando un sistema de intercambio de transcripciones de muestras de habla entre los investigadores de todo el mundo, se gestó desde 1981 hasta 1984, en que se puso en marcha, el Proyecto CHILDES (*CHILd Language Data Exchange System*), coordinado por Brian MacWhinney y Catherine Snow (MacWhinney y Snow, 1985, 1990), cuyos objetivos enuncia MacWhinney (1995, p. X) así:

1. Proporcionar más datos de más niños hablantes de más lenguas.
2. Obtener mejores datos mediante un sistema de transcripción consistente y documentado.
3. Automatizar el proceso de análisis de los datos.

Herramientas del sistema CHILDES

El proyecto encaró dichos objetivos desarrollando tres herramientas distintas e integradas en su función, «como las patas de un taburete», tal y como lo expresa MacWhinney (1995) en el libro que constituye el *Manual* del sistema CHILDES:

1. *DATABASE*. El objetivo fundamental del proyecto era crear una base de datos informatizada con el conjunto de todas las transcripciones que los investigadores miembros del sistema fueran aportando, esta es pues la primera herramienta, la de los datos. Una vez informatizadas las muestras fundacionales de Brown (Adán, Eva y Sara), constantemente se ha ido actualizando *DATABASE* con nuevas transcripciones en cada uno de sus apartados: *inglesas, no-inglesas, bilingües, narrativas y con alteraciones del lenguaje*, hasta alcanzar unas dimensiones considerables tanto cuantitativas, como cualitativas, es decir, en número de lenguas representadas. Sin embargo, todavía son mayoritarias las transcripciones en inglés y llama la atención la escasez de

datos en español, que se reducen a seis *corpora*, en relación con la importancia y difusión de esta lengua, por lo que uno de los objetivos del presente artículo es promover el enriquecimiento de DATABASE con nuevos datos en español.

2. *CHAT* (Codes for the Human Analysis of Transcripts). El primer problema al que se enfrentaba el proyecto de intercambio era la ausencia de sistematización en la forma tradicional de las transcripciones (Ochs, 1979). Un requisito básico era pues establecer una herramienta de transcripción, un sistema estandarizado, un formato, el formato CHAT, que han de tener las transcripciones para entrar a formar parte de DATABASE. Existen unos mínimos requerimientos formales que se incluyen dentro de lo que se considera la codificación más elemental, *MinCHAT*, que luego expondremos, y la posibilidad de codificaciones más complejas con finalidades investigadoras específicas.

3. *CLAN* (Computerized Language Analysis). La informatización de las transcripciones facilitó enormemente lo que al fin y al cabo constituye el objetivo último de la metodología RETAMHE, el análisis de los resultados obtenidos en el proceso de registro y codificación, tarea de recuento que antaño se hacía «a mano» en sesiones agotadoras y con fiabilidad excesivamente dependiente de los errores del investigador. La automatización de dicha tarea se concretó en la tercera herramienta, CLAN, un paquete de programas informáticos específicos para el análisis de transcripciones en CHAT. Al igual que el sistema de codificación, los programas tienen un uso básico, *minCLAN*, que introduciremos en este artículo, y unos usos avanzados que se explican pormenorizadamente en el Manual (MacWhinney, 1995).

El sistema incluye versiones tanto para Macintosh, como para MS-DOS y recientemente se ha incorporado una nueva para Windows 95. Las actualizaciones de todos los componentes se publican anualmente en

el CD-ROM «The CHILDES Database» (MacWhinney, 1997) que incluye las tres herramientas citadas, además del manual y de una cuarta herramienta la *BIB/DATABASE*, compendio bibliográfico actualizado sobre el lenguaje infantil con cerca de 13.000 entradas (Higginson y MacWhinney, 1990).

Codificación básica de transcripciones (MinCHAT)

Las transcripciones en CHAT no presentan inicialmente el formato tradicional en columnas, sino en líneas y se escriben con el procesador de textos de CHILDES (CED) (véase un ejemplo en la Tabla 1). Existen tres tipos de líneas, indicado cada uno de ellos por un símbolo de codificación que aparece como primer carácter de la línea:

1. Líneas de encabezamiento @ Ej. @Begin
2. Líneas principales * Ej. *CHI:
3. Líneas dependientes % Ej. %err:

Líneas de encabezamiento (@)

Como su nombre indica, son en su mayoría las líneas que encabezan la transcripción, marcan el inicio y el final de la misma o pasajes determinados y contienen datos fundamentales tales como quiénes son los participantes en la conversación, la edad del niño objeto de estudio, la fecha de la grabación o la situación en que se realizó la misma. Se clasifican en tres categorías: obligatorias, constantes y variables.

1. Obligatorias

@Begin: Primera línea de una transcripción

@End: Última línea de una transcripción

@Participants: Segunda línea de una transcripción donde se listan todos los participantes con las tres letras que identificarán sus intervenciones (CHI, INV, MOT, etc.),

con su nombre (opcional) y su papel en la situación (Target_Child, Investigator, Mother, etc.)

Ej. @Begin
@Participants: CHI Sara
Target_Child, INV Gerardo Investigator
**INV: hola # Sara!*
**CHI: hola!*
@End

2. Constantes

@Age of XXX: Edad del participante XXX (años;meses.días)

Ej. @Age of CHI: 4;1.3

@Birth of XXX: Fecha de nacimiento del participante XXX (día-MES-año)

Ej. @Birth of CHI: 23-DEC-1993

@Coder: Identificación del transcriptor/es

Ej. @Coder: Verónica Martínez López

@Education of XXX: Años de escolaridad del participante XXX (0-20)

@Filename: Nombre del archivo informático que contiene la transcripción

Ej. @Filename: angelpcp.cha

@Language: Lengua principal de los participantes en la transcripción

Ej. @Language: Spanish

@SES of XXX: Status socioeconómico del participante XXX

Ej. @SES of CHI: middle

@Sex of XXX: Sexo del participante XXX (female/male)

3. Variables

@Date: Fecha de la grabación (día-MES-año)

Ej. @Date: 26-JAN-1998

@Location: Lugar de la grabación (ciudad, provincia, país)

Ej. @Location: Oviedo,Asturias,Spain

@Time Duration: Horas de inicio y final del registro (00:00-24:00)

Ej. @Time Duration: 18:45-19:25

@Situation: Situación en la que tiene lugar la grabación

Ej. @Situation: en el salón del domicilio del niño, sentados sobre la alfombra

@Activities: Actividades de los participantes durante la grabación

Ej. @Activities: jugando con puzzles y mirando cuentos

@Comment: Comentario sobre algún aspecto global de la grabación

Ej. @Comment: el niño estuvo inquieto porque era su hora de la siesta

@New Episode: Marca de un nuevo episodio cuando la grabación no se hace en una única sesión o hay interrupciones o modificaciones en la situación.

@Bg: comienzo de un fragmento (gem) que podrá ser analizado aisladamente por los programas de CLAN.

Ej. @Bg: cuento de Caperucita

@Eg: final de un fragmento «gem»

Ej. @Eg: cuento de Caperucita

Líneas principales (*)

Son las líneas de la transcripción propiamente dichas donde se recoge todo lo que dicen los participantes. Comienzan con el símbolo * y las tres mayúsculas identificativas del participante seguidas de dos puntos. El texto o transcripción de lo hablado empieza en la novena columna con lo que se debe introducir el espacio correspondiente mediante un tabulador prefijado en CED, el procesador de textos de CHILDES. No se utilizan mayúsculas excepto para los nombres propios. El final de cada línea principal debe delimitarse siempre, bien con un punto, una interrogación o una exclamación, con lo que dichos símbolos no pueden aparecer en otra posición que no sea el final de una línea principal. Cada línea no necesariamente recoge un turno completo del participante, sino un único enunciado completo,

siendo necesario utilizar una nueva línea por cada nuevo enunciado, aunque el participante sea el mismo y no cambie el turno. Por ejemplo, cuando el niño narra un cuento, probablemente será necesario incluir lo que dice en varias líneas principales:

*CHI: *la mamá le dijo que llevase la cestita.*

*CHI: *y entonces Capercucita se metió en el bosque.*

1. Códigos en líneas principales

Pausas

Sólo excepcionalmente se emplean comas, hasta el punto de que puede ser preferible no utilizarlas nunca. Al tratarse del habla, lo que se marca son las pausas reales de los hablantes y se indican mediante el símbolo: #

*CHI: *no # dame el coche # es mío.*

*CHI: *no # dame # el coche es mío.*

*CHI: *no # dame # el coche # es mío.*

Material ininteligible

xxx fragmento que no se entiende

*CHI: *xxx.*

%exp: *pasa un coche y no se oye lo que dice.*

xx palabra/s que no se entiende/n

*CHI: *no # dame # el coche # xx xx.*

[?] palabra o fragmento dudoso

*CHI: *no # dame # el coche # <es mío> [?].*

www material no transcrito

*CHI: *www.*

%exp: *el niño canta una canción*

*INV: *www.*

%exp: *habla con el operador de la cámara*

& palabra incompleta o no palabra

*CHI: *&ee no # dame # el &co coche # &secí [?].*

Acciones paralingüísticas

[=!] acciones que acompañan o sustituyen al habla y tienen un valor comunicativo

*CHI: *<me quitó el camión> [=! grita].*

*CHI: *0 [=! ríe].*

*INV: *0 [=! asiente].*

El 0 indica que en su turno el niño no dice nada, pero ejecuta alguna acción que puede ser paralingüística o no:

*INV: *cómo te llamas?*

*CHI: *0 [% la niña mira a la cámara].*

Interjecciones

Van seguidas de una pausa que sustituye al signo exclamativo, salvo cuando van al final de la línea principal.

*INV: *ah # sí!*

Onomatopeyas y palabras compuestas

Cuando deseamos que una palabra o conjunto de palabras se compute como una sola, como puede ser el caso de las onomatopeyas o los nombres compuestos, se unen mediante el símbolo: +

brn+brn+brn/ja+ja+ja

taca+taca/guau+guau

Jose+Luis/Caperucita+Roja/Power+Ranger

(el símbolo + puede sustituir también a los guiones, ya que estos son interpretados como códigos morfológicos por CLAN).

Tabla 1

Fragmento de transcripción en formato CHAT de una muestra de habla espontánea recogida en la conversación entre un niño de tres años y su madre

@Begin
 @Participants: CHI Target_Child, MOT Mother
 @Age of CHI: 3;6.0
 @Coder: Eliseo Diez-Itza
 @SES of CHI: middle
 @Sex of CHI: male
 @Filename: angelpcp.cha
 @Situation: la madre y el niño juegan sobre la alfombra del salón sentados frente a frente. Tienen un papel y una caja con pinturas y sellos de caucho.
 @Comment: fragmento de una transcripción con finalidad didáctica
 @Language: Spanish
 @Location: Gijón, Asturias, Spain

*CHI: Ángel # Pepe # y Luis.
 %act: va señalando alternativamente los dibujos en el papel
 *MOT: ah # Angel Pepe y Luis?
 *MOT: andá # el tractor de quién es # este?
 *CHI: de Luis.
 *MOT: y este # qué?
 *MOT: u:y # venga!
 *CHI: este de Angel.
 *MOT: ahora pintamos # por ejemplo +...
 *MOT: a quién ponemos aquí?
 *CHI: eh [?]?
 %com: el niño emite un sonido poco audible como si no hubiese entendido bien
 *MOT: <a Ana> [/] pintamos a Ana # eh?
 *MOT: venga # vamos a pintar a Ana.
 *CHI: no # esta es Josefa.
 *MOT: esta es Josefa?
 %act: señala un dibujo en el papel
 *MOT: venga # yo te la pongo ahí # y tú le das colores # vale?
 *MOT: de qué color vas a pintar a Josefa?
 *CHI: <tu me lo> [/] tu moja [= sello de caucho] # y lo pones aquí.
 %act: señala un lugar en el papel
 *MOT: yo moja y lo pongo aquí.
 %act: la madre coge un sello de caucho y lo imprime donde ha señalado el niño y tú qué haces?
 *CHI: <yo # nada> [?] +...
 %com: parece rehuir la pregunta y a continuación cambia de tema
 *CHI: todo esto lo pinté yo y esto lo pintó xx xx xx.
 %act: se pone a señalar dibujos en el papel, sin levantar la vista
 CHI: mira # como come hierba el burrito # se ponió [] verde!
 %err: ponió=puso \$MOR;
 *MOT: no me digas # se puso verde porque comió hierba?
 *MOT: a ver # entonces # Josefa [/] vamos a pintar a Josefa.
 *MOT: mira # porque lleva ahí la cesta de qué?
 *CHI: de Caperucita.
 *MOT: de Caperucita?
 *MOT: ah # yo pensé que era de ir a por los huevos a las gallinas.
 *MOT: venga!
 @End

Interrupciones

+/. final de una producción incompleta por haber sido interrumpida

*INV: *vamos a ir +/.*

*CHI: *al cine!*

+... final de una producción incompleta sin interrupción

*CHI: *pues el otro día +...*

*CHI: *a mí me gusta mucho la playa.*

+//. final de una producción autointerrumpida

*INV: *no quiero que os +//.*

*INV: *mira # aquí viene mamá!*

+/? Final de una pregunta interrumpida

*INV: *queréis ir al +/?*

Repeticiones, reformulaciones y superposiciones

[/] repetición exacta

*CHI: *<tú moja> [/] tú moja y lo pones aquí.*

[//] reformulación: sobre la misma idea cambia alguna palabra o la forma de la frase

*CHI: *<tú me lo> [//] tú moja!*

[-/] falso comienzo sin reformulación posterior, es decir, empezando una frase nueva

*CHI: *<tú me lo> [-/] # dame esa pintura!*

[>] superposición con la línea siguiente

[<] superposición con la línea anterior

*INV: *yo lo pongo ahí # <y tú le das colores> [>].*

*CHI: *<tú moja y lo> [<] pones ahí!*

2. Explicaciones en líneas principales

[= texto] explicación que identifica personas o cosas

*INV: *dáselo [= lápiz] a ella [= Sara].*

[:=x texto] traducción

*CHI: *vamos con lololo@c [:=x abuelo]!*

(@c colocado al final de la palabra «lololo» indica que se trata de un término inventado por el niño, que forma parte de su jerga individual)

[=? texto] transcripción alternativa

*CHI: *dame # el &co coche # &secí [=? sadí].*

[% texto] comentario

*INV: *y tú # qué haces?*

*CHI: *<yo # nada> [% en tono evasivo]*

Líneas dependientes (%)

Contienen códigos, comentarios o descripciones que complementan la transcripción propiamente dicha que aparece en las líneas principales de las que dependen. Comienzan con el símbolo % seguido de tres letras minúsculas identificativas de su nombre/contenido y dos puntos, no llevando puntuación al final excepto en el caso de las líneas de error en las que se emplea el punto y coma. Se colocan siempre a continuación de la línea principal de la que dependen, es decir, a la que se refieren, y se refieren sólo a ella. Pueden seguir a una línea principal tantas líneas dependientes como se considere necesario introducir, pero no se pueden repetir sus nombres, de modo que, por ejemplo, no podemos codificar varios errores que aparezcan en la misma línea principal mediante líneas %err:, hemos de incluirlos todos en una misma línea con ese nombre y separados por punto y coma. Los programas de CLAN reconocen 23 nombres de líneas dependientes llamadas «standard», que cumplen distintas finalidades, pero se pueden inventar nuevas líneas en el caso (raro) de que aquellas no cubran nuestras necesidades. Veamos algunas de ellas:

1. Líneas de información y observaciones

%act: acciones de los participantes

%com: comentarios
%exp: explicaciones
%par: material paralingüístico
%sit: información situacional
%flo: versión fluida o «limpia» de la transcripción

**INV*: yo [/] yo lo pongo [= un sello de caucho] ahí [= en el papel] # <y [/] # pero tú le das colores> [>].

%flo: yo yo lo pongo ahí y pero tú le das colores.

**CHI*: <tú moja [= el sello de caucho] y lo> [<] pone [*] ahí!

%flo: tú moja y lo pone ahí.

2. Líneas de codificación

%cod: línea de codificación general

**CHI*: se lo di a él.

%cod: \$MLU=5 \$PRON=3 \$PREP=1 \$VERB=1

%pho: descripción codificada de la fonología (cfr. cap. 11 (PHONASCH) y 10 (UNIBET) del Manual)

%spa: actos de habla (cfr. cap. 13 del Manual)

**INV*: hola # Sara!

%spa: \$nma:cl

**CHI*: hola!

%spa: \$sat:ac

(códigos INCA-A: Ninio, Snow, Pan y Rollins, 1994).

%mor: descripción codificada de la morfosintaxis (cfr. cap. 14 del Manual)

**CHI*: se cayó el coche.

%mor: pro:refl/se v/caer-PAST det/el n/coche

%err: exige que en la línea principal los errores estén marcados con [*], finaliza con punto y coma (cfr. cap. 12 del Manual)

**CHI*: mira # es un pínsite [*]!

%err: pínsite = príncipe \$PHO;

**CHI*: lo ponió [*] aquí.

%err: ponió = puso \$MOR;

**CHI*: <me se> [*] cayó!

%err: me se = se me \$SYN;

**CHI*: un caballo [*].

%exp: señala un dibujo de una vaca

%err: caballo = vaca \$LEX;

**CHI*: dame el coche # &secí [*] [?].

%err: secí = ?; \$PITO.

**CHI*: pínsite [*] ponió [*] gorro> [*].

%err: pínsite = el príncipe \$PHO \$SYN; ponió = se puso \$MOR \$SYN; gorro = la orona \$LEX; \$SYN.

Análisis computerizado de transcripciones (MinCLAN)

Una vez transcritos los datos, CHILDES proporciona la posibilidad de automatizar el tercer y último proceso de la metodología RETAMHE, el análisis. Ello merced al paquete de programas informáticos CLAN que Leonid Spektor diseñó en la Universidad Carnegie Mellon. Estos programas permiten realizar recuentos de frecuencias, cálculos del MLU, búsqueda de palabras, análisis de coocurrencias, de interacciones, de cambios textuales, etc. Su formato se adapta al de las transcripciones en CHAT, aunque muchos de ellos funcionan en archivos ASCII.

El núcleo de los comandos y opciones de CLAN, constituye el sistema MinCLAN, cuya utilización requiere únicamente conocimientos básicos de los sistemas operativos de Macintosh o PC y aprendizaje previo del sistema MinCHAT que hemos resumido en el apartado anterior de transcripción usando el procesador de textos de CHILDES (CED). En el Manual se ofrece una descripción detallada de los programas y sus múltiples opciones, empezando según un orden alfabético precisamente con CED (MacWhinney, 1995, cap. 21). Las opciones de cada programa pueden consultarse también di-

rectamente en la pantalla del ordenador simplemente escribiendo el nombre del programa y pulsando «enter», veamos algunos de ellos.

MLU y MLT

La longitud media de los enunciados ha sido la medida más universalmente empleada para cuantificar el desarrollo gramatical de los niños. Naturalmente CLAN proporciona un programa llamado MLU para efectuar dicho cálculo sobre una transcripción o un conjunto de ellas. Sin embargo, y particularmente para el español, resulta difícil establecer dicha medición en términos de morfemas, por lo que concedemos mayor fiabilidad al cálculo del MLU en función del número de palabras por enunciado (Diez-Itza, 1992). Dicha medición es uno de los resultados de otro programa que mide la longitud media de los turnos: MLT. Dicho programa, como se puede ver en el ejemplo de la Tabla 2, donde se aplica al fragmento de transcripción reproducido arriba, mide para cada uno de los participantes el número de enunciados, el número de turnos y el número de palabras y sobre ellos calcula la media de palabras por turno, la media de producciones por turno y la media de palabras por enunciado, es decir, el MLU, la longitud media de los enunciados. Para ejecutar los programas de CLAN basta con teclear su nombre seguido del nombre del archivo que contiene la transcripción que deseamos analizar. Por ejemplo, para obtener en pantalla el análisis que aparece en la Tabla 2 el comando es:

>mlt angelpep.cha

FREQ

Este es uno de los más poderosos programas de CLAN y a la vez uno de los más fáciles de usar, por lo que se recomienda

iniciarse en MinCLAN a través de sus distintos usos. *FREQ* efectúa sobre las transcripciones recuentos de las frecuencias de todas las palabras ofreciendo un listado de las mismas, al tiempo que calcula otro de los índices tradicionalmente más utilizados para estudiar el vocabulario la *type/token ratio* (TTR), índice de diversidad léxica que viene dado por el cociente entre los vocablos diferentes que aparecen en la muestra y el total de palabras de la misma. La ejecución de *FREQ* se realiza del mismo modo señalado para MLT y MLU, sin embargo el output que se obtiene no separa las frecuencias de cada uno de los participantes, es decir, nos da el recuento de todas las palabras que aparecen en la transcripción. Puesto que normalmente nos interesará realizar el recuento de un participante, es necesario introducir una opción en el comando de modo que el análisis se reduzca a un tipo de línea, dicha opción es *+t* seguido del nombre de la línea que deseamos anali-

<i>Tabla 2</i> Análisis con MLT de la longitud media de los turnos y enunciados del niño y de la madre en la transcripción de la Tabla 1
<pre> MLT.EXE angelpep.cha +f Sat Apr 11 21:03:01 1998 MLT.EXE (18-JUL-95) is conducting analyses on: ALL speaker tiers ***** From file <angelpep.cha> to file <angelpep.mt> MLT for Speaker: *MOT: MLT (xxx and yyy are INCLUDED in the utterance and morpheme counts): Number of: utterances = 19, turns = 8, words = 110 Ratio of words over turns = 13.750 Ratio of utterances over turns = 2.375 Ratio of words over utterances = 5.789 MLT for Speaker: *CHI: MLT (xxx and yyy are INCLUDED in the utterance and morpheme counts): Number of: utterances = 10, turns = 8, words = 45 Ratio of words over turns = 5.625 Ratio of utterances over turns = 1.250 Ratio of words over utterances = 4.500 </pre>

zar: Por ejemplo, para obtener en pantalla un análisis de frecuencia léxica del niño (*CHI) como el del ejemplo de la Tabla 3, realizado sobre el fragmento de la transcripción reproducida más arriba la orden sería:

```
>freq +t*chi angelpép.cha
```

Nótese que el recuento total de palabras de *FREQ* no coincide con el de *MLT*. Ello se debe a que *FREQ* cuenta por defecto todas las palabras, mientras que *MLT* excluye por defecto las repeticiones y reformulaciones que estén codificadas como tales. Si deseamos que *FREQ* excluya dichas palabras hemos de añadir la opción *+r6*.

<i>Tabla 3</i>	
Análisis con <i>FREQ</i> de la diversidad léxica del niño en la transcripción de la Tabla 1	
<pre>FREQ.EXE +t*chi angelpép.cha +f Sat Apr 11 21:03:45 1998 FREQ.EXE (18-JUL-95) is conducting analyses on: ONLY speaker main tiers matching: *CHI; ***** From file <angelpép.cha> to file <angelpép.frq></pre>	
<pre>1 aquí 1 burrito 1 caperucita 1 come 1 como 3 de 1 eh 1 el 1 es 1 esta 1 este 2 esto 1 hierba 1 josefa 4 lo 2 luis 1 me</pre>	<pre>1 mira 1 moja 1 nada 1 no 1 pepe 1 pinté 1 pintó 1 pones 1 ponió 1 se 1 todo 2 tu 1 verde 3 xx 3 y 2 yo 2 Ángel</pre>
<pre>34 Total number of different word types used 48 Total number of words (tokens) 0.708 Type/Token ratio</pre>	

KWAL

Este programa busca las producciones o líneas de la transcripción en las que aparece determinada palabra que se especifique. Permite por tanto analizar el contexto en el que aparecen elementos predeterminados. Dichos elementos se señalan mediante la opción *+s* seguida de la palabra o grupo de palabras en cuestión entrecomillados. También podemos escribir en un archivo una lista de las palabras que deseamos buscar, como artículos, preposiciones, etc. (@art, @prep, etc.) y *KWAL* las buscará en la transcripción:

```
>kwal +t*chi +s«en» angelpép.cha
>kwal +t*chi +s@prep angelpép.cha
```

COMBO

Es también un programa de búsqueda con *+s*, pero en este caso de cadenas de caracteres, palabras o grupos de palabras que pueden aparecer o no juntas en la transcripción. Los elementos de la cadena que se desea buscar deben unirse en el comando mediante el símbolo *^*. Este programa resulta particularmente importante para la investigación de aspectos sintácticos. Podemos buscar, por ejemplo, la secuencia «come burrito» en una producción, con la

<i>Tabla 4</i>	
Análisis con <i>COMBO</i> de la aparición de la secuencia «come^*burrito» en la transcripción de la Tabla 1	
<pre>COMBO.EXE +scome^*burrito angelpép.cha Sun Apr 12 12:26:19 1998 COMBO.EXE (18-JUL-95) is conducting analyses on: ALL speaker tiers ***** From file <angelpép.cha></pre>	
<pre>*** line 45; file angelpép.cha *** *CHI: mira # como come hierba el burrito # se ponió [*] verde! 1 1 Strings matched 1 times</pre>	

opción de que dichas palabras puedan no ser consecutivas en la misma, en cuyo caso insertaremos en el comando el símbolo * (el resultado de este análisis sobre la transcripción del ejemplo es la que aparece en la Tabla 4):

```
>combo +s«come^*^burrito» angelpep.cha
```

CHIP

Se trata de un programa escrito por Jeffrey Sokolov para estudiar la interacción verbal, es decir, para analizar cuantitativamente las similitudes y diferencias entre el input lingüístico parental y las respuestas infantiles. CHIP compara dos grupos de enunciados predeterminados, los primeros a los que llamamos «fuentes» y marcamos con +b y los segundos a los que llamamos «respuestas» y marcamos con +c. El programa introduce líneas de codificación con las distintas comparaciones y un cuadro final donde se recogen las estadísticas de los análisis. Por ejemplo, con la siguiente orden obtenemos un output con fragmentos del tipo de los que aparecen en la Tabla 5:

```
>chip +bMOT +cCHI angelpep.cha
```

Tabla 5
Análisis comparativo con CHIP de enunciados «fuente» (madre) y enunciados «respuesta» de la transcripción de la Tabla 1
<pre>CHIP.EXE +bMOT +cCHI angelpep.cha +f Sat Apr 11 21:27:12 1998 CHIP.EXE (18-JUL-95) is conducting analyses on: ALL speaker tiers ***** From file <angelpep.cha> to file <angelpep.chp> *CHI: Ángel # Pepe # y Luis . *MOT:ah # Ángel Pepe y Luis ? %adu: \$EXA:Ángel-pepe-y-luis \$ADD:ah \$EXPAN \$DIST = 1 \$REP = 0.80 *MOT:andá # el tractor de quién es # este ? %asr: \$NO_REP \$REP = 0.00 %adu: \$NO_REP \$REP = 0.00 [...]</pre>

GEM

Cualquiera de los programas de CLAN puede aplicarse a un fragmento de la transcripción predeterminado, sin necesidad de copiarlo en un archivo separado. GEM es el programa que separa dicho fragmento y se utiliza en combinación con los otros programas. El fragmento o «gem» en cuestión debe estar delimitado por las líneas de encabezamiento antes citadas @Bg y @Eg. Puede interesarnos, por ejemplo, analizar con FREQ la diversidad léxica en una tarea narrativa, como contar el cuento de Caperucita Roja, que está incluida en una transcripción con las marcas de «gem» correspondientes, para lo que utilizaríamos GEM con el siguiente comando:

```
>gem +scaperucita +d angelpep.cha /
freq +t*chi
```

GEM corta o separa ese fragmento para que FREQ o cualquier otro programa realice el análisis solamente sobre el mismo, permitiendo así obtener medidas de diferentes tareas o situaciones que se pueden comparar entre sí, con otras similares o con el cómputo total de la transcripción en la medida correspondiente.

Normas elementales para la utilización de CHILDES

Existe ya una amplia literatura de investigaciones que han utilizado o se han basado en el sistema CHILDES y un número menor, aunque creciente, de trabajos, como el presente, que tratan de difundir, enseñar o promover su empleo. De entre estos destaca un Manual específicamente dedicado a prácticas de investigación del desarrollo del lenguaje utilizando CHILDES (Sokolov y Snow, 1994). En él se recogen ejemplos de investigaciones de temática diversa, desde el «babytalk» a las diferencias individuales

o desde los artículos en español al perfil de los niños con retraso específico del lenguaje. Aparece, cómo no, un capítulo donde se analizan las transcripciones de Brown, la de Adán en concreto, y un simpático y breve prólogo de Brown que nos devuelve al inicio de este artículo con una anécdota metodológica: dos investigadores japoneses preguntaron a Bellugi y Brown, por el laboratorio y el equipo con el que habían estudiado a Adán, Eva y Sara, «ambos empuñamos en alto nuestros lápices amarillos Mogol No. 2» (Sokolov y Snow, 1994, p. iX).

Seguimos empuñando metafóricamente esos lápices de nuestros antepasados cuando empleamos la metodología RETAMHE y las herramientas que proporciona el proyecto CHILDES. Animamos pues a los investigadores del comportamiento lingüístico y su desarrollo en el ámbito del español a seguir esta senda o método. Tal y como MacWhinney indica en el *Manual* y también en el *00Readme* que acompaña al CHILDES CD-ROM, las personas que utilicen los datos y los programas de CHILDES deben seguir las siguientes normas:

1. Hacerse miembro del sistema CHILDES poniéndose en contacto a través de la dirección electrónica: childes@andrew.cmu.edu, facilitando nombre, dirección, e-mail, teléfono y explicando el posible uso que desea hacer de los datos y los programas.

2. Adquirir una copia del *Manual* (MacWhinney, 1995) y de los programas

a la editorial Erlbaum. El manejo correcto de los datos requiere una comprensión básica de los contenidos del mismo, incluyendo el sistema CHAT, los programas CLAN y la documentación descriptiva de DATABASE.

3. Citar a MacWhinney (1995) en cualquier publicación que se base en el uso de los datos y los programas, así como los artículos vinculados a los *corpora* particulares que se manejen.

4. Asumir la responsabilidad de contribuir al crecimiento de la base de datos aportando nuevos *corpora* y urgiendo a otros colegas a que hagan lo mismo.

¡Buena suerte!

Agradecimientos

Este trabajo está dedicado a la memoria de Roger Brown, uno de los pioneros en hacer realidad el sueño de compartir datos del lenguaje infantil. Está relacionado con diversos cursos y seminarios metodológicos impartidos por los autores en España, de modo que desean dar testimonio de gratitud a los participantes y organizadores de dichos eventos por su contribución a la difusión del proyecto. Muy especialmente deben agradecer la inestimable colaboración en los mismos de Verónica Martínez, de la Universidad de Oviedo, y María Carrasco y Cruz Celis, de la Universidad Complutense de Madrid, autoras de distintos documentos y manuales de gran utilidad acerca de CHILDES.

Referencias

Arnau, J. y Boada, H. (1985). Aspectos de la evolución del instrumento lingüístico en bilingües y monolingües. *III Congreso Nacional de AESLA*. Valencia.

Brown, R. (1973). *A first language. The early stages*. Cambridge, Mass.: Harvard University Press.

Bruner, J. (1986). *El habla del niño*. Barcelona: Paidós. (Orig. 1983).

Diez-Itza, E. (1992). *Adquisición del lenguaje*. Oviedo: Pentalfa.

Higginson, R. y MacWhinney, B. (1990). *CHILDES/BIB: An annotated bibliography of child language and language disorders*. Hillsdale, NJ: Erlbaum.

MacWhinney, B. (1995) *The CHILDES Project. Tools for analyzing talk*. Hillsdale, N.J.: Erlbaum.

MacWhinney, B. y Snow, C. (1985). The Child Language Data Exchange System. *Journal of Child Language*, 12, 271-296.

MacWhinney, B. y Snow, C. (1990). The Child Language Data Exchange System: an update. *Journal of Child Language*, 17, 457-472.

Ninio, A., Snow, C., Pan, B. y Rollins, P. (1994). Classifying communicative acts in children's interactions. *Journal of Communicative Disorders*, 27, 157-188.

Ochs, E. (1979). Transcription as theory. En E. Ochs y B. Schieffelin (eds.) *Developmental Pragmatics* (pp. 43-72). Nueva York: Academic Press.

Serra, M. (1976). Métodos e índices para el estudio psicológico y patológico del lenguaje. *Anuario de Psicología*, 15: 171-196.

Siguán, M. (1983). *Metodología para el estudio del lenguaje en la infancia*. Barcelona: Publicacions i Edicions de la Universitat.

Siguán, M., Colomina, R. y Vila, I. (1990). *Metodología para el estudio del lenguaje Infantil*. Vic: Abril.

Sokolov, J.L. y Snow, C. (1994). *Handbook of research in language development using CHILDES*. Hillsdale, N.J.: Erlbaum.

Solé, M.R. (1988). *La comunicació verbal en el marc escolar*. Vic: Eumo.

Triadó, C. y Forns, M. (1989) *La evaluación del lenguaje: Una aproximación evolutiva*. Barcelona: Anthropos.

Acceptado el 23 de diciembre de 1998