TRANSCRIPCIÓN, SEGMENTACIÓN E INFORMATIZACIÓN DE UN CORPUS DE LENGUAJE

Por **J. A. Rondal, J. F. Bachelet, C. Grana Monterin, A. M. Le Docte, F. Peree** y **E. Stafilas**

El trabajo del lingüista, del sociolingüista y del psicolingüista implican a menudo el análisis e interpretación de un corpus de lenguaje. Por «corpus de lenguaje» (hoy se habla cada vez más de «texto» —véase Denhière, Thibaut y Rondal, 1984— pero en este artículo conservaremos la denominación de «corpus»), debe entenderse todas aquellas producciones verbales cuya extensión supere la frase, proveniente de sólo uno o de varios hablantes en una conversación, y analizables en enunciados más elementales. La interpretación de los datos lingüísticos recogidos y analizados depende, por supuesto, de los objetivos inmediatos de la investigación y de las referencias teóricas del investigador. Este artículo no tratará de aquellos problemas. El análisis de un corpus de lenguaje, desde luego, nunca es inocente. Se encuentra determinado en gran medida por las referencias teóricas del investigador y por los objetivos que éste se ha fijado en su investigación. Tanto la recolección como el primer tratamiento de un corpus de lenguaje son más estandarizables.

En este plano, el problema es la ausencia relativa de indicadores metodológicos precisos que constituyan referencias para la práctica. La situación es insatisfactoria puesto que complica la tarea de los investigadores, sobre todo si se trata de principiantes. Además, no existe ninguna seguridad de que todos los investigadores en este campo utilicen aproximadamente la misma metodología para recoger y analizar los datos lingüísticos sobre los cuales más tarde fundarán sus análisis y sus interpretaciones.

No sabemos de ningún documento detallado sobre las reglas básicas que deben seguirse en francés para recoger y, sobre todo, segmentar y transcribir un corpus de lenguaje. En inglés podremos encontrar algunas indicaciones, aunque bastante insuficientes, en Brown (1973).

Con el fin de superar esta laguna, al menos en parte, y de responder a las preguntas que regularmente nos plantean sobre estos puntos, hemos decidido redactar y publicar el siguiente documento. No nos arrogamos, desde luego, ninguna pretensión de exhaustividad ni perseguimos la creación de un monopolio metodológico. Su fin consiste en dar fiel cuenta de los aspectos esenciales de nuestros procedimientos durante un estudio longitudinal de gran envergadura, enfocado sobre las interacciones verbales madre-hijo en la adquisición del lenguaje materno (en este caso, el francés) por parte del niño.

Este estudio ha permitido recoger unas cien horas de grabación del lenguaje intercambiado en una situación natural de «juego libre» entre una madre y su hijo, durante intervalos de tiempos regulares entre los dos y cinco años aproximadamente. Este enorme corpus conversacional debió ser sometido, en un primer momento, a una completa transcripción y luego a una segmentación en enunciados.

Debido a que este artículo ha sido escrito originariamente en francés algunos ejemplos no se han traducido porque podía modificarse de forma importante el sentido dado por el autor. Solamente se han traducido aquellos en los que la estructura dada como ejemplo tenía un equivalente muy aproximado en castellano.

Correspondencia: *Laboratorio de Psicología del Lenguaje de Ciencias de la Educación. Universidad de Lieja, B. 32. Sart y Servicio de matemáticas aplicadas. Instituto de Psicología y Tilman 4000 Lieja.*

Una vez hecho esto, debían dictarse reglas muy precisas de transcripción y de segmentación. Un trabajo tan simple en apariencia es, en realidad, muy difícil y complicado, como ya lo saben los que han emprendido este tipo de estudios. Pensamos que las reglas y los principios metodológicos a los que llegamos pueden ser utilizados en otras investigaciones, y ésta es la razón de ser del presente documento.

La transcripción y segmentación de un corpus de lenguaje no tiene otro objetivo que el permitir un posterior análisis. Éste depende, claro está, del contexto teórico en el que se inscribe la investigación y de los presupuestos metodológicos del investigador. Los nuestros se acercan a lo que podríamos denominar la perspectiva *medioambientalista* de la construcción del lenguaje en el niño, tal como ha sido definido y documentado por Ronda1 (1981, 1983) y por Moerk (1983).

Cuando los datos que inspiraron el presente artículo hayan sido completamente analizados e interpretados, podrán constituir el objeto de un informe detallado y completo del procedimiento analítico utilizado.

Los elementos de análisis extraídos del presente trabajo son simplemente ilustrativos de la técnica de la informática, a la cual estimamos útil recurrir, dadas las dimensiones del corpus. También en este plano pensamos que el trabajo realizado para informatizar una parte de nuestros análisis puede representar algún interés para nuestros colegas, razón por la cual lo incluimos en este informe. Sin embargo, los elementos de análisis de los que daremos cuenta no representan más que una muy pequeña parte del trabajo (y no la más importante) referida al cálculo de frecuencias distributivas de una serie de clases y subclases formales en el corpus, organizadas según sus diferentes subdivisiones temporales. Del mismo modo, y con un fin ilustrativo, se hace mención de algunos índices conversacionales.

Se pueden calcular los diversos índices en el ordenador con la condición de que se utilicen los programas adecuados, principio que hemos cumplido. En la última parte del artículo proporcionamos las indicaciones que se refieren a esta etapa del análisis.

El estado actual de la técnica informática no permite que la parte más importante de nuestro trabajo analítico, dadas sus orientaciones teóricas, sea informatizada (Rondal, 1983) (de hecho, se puede lograr en teoría, pero su realización sería poco rentable, dada la complejidad de los programas que habría que producir). Esto no disminuye en nada el interés que puede representar la informatización de un corpus lingüístico con el fin de realizar automáticamente una serie de análisis de «primer nivel», lo que permitiría extraer una «radiografía general» (que aún sigue siendo muy aproximativa) del corpus (por ejemplo, en este caso, la evolución de la extensión media de los enunciados de la madre y del niño, desde el principio hasta el final del corpus, o la frecuencia distributiva de las diferentes clases y subclases formales según la evolución temporal del corpus, etc.).

Llegado este punto, cabe tratar la exposición técnica propiamente dicha. Como se ha indicado más arriba, el análisis del lenguaje suscita el problema de su transcripción, de su segmentación y, eventualmente, de su informatización.

La transcripción: el primer paso

De ordinario, la lengua escrita satisface sólo en parte la exigencia de correspondencia biunívoca entre sonido y grafía. El Alfabeto Fonético Internacional (AFI), como modificación de la lengua escrita, es preferible a cualquier otro sistema de notación que establezca una correspondencia satisfactoria entre los sonidos y los símbolos gráficos. Si un estudio trata, única o parcialmente, de la pronunciación de los sonidos, uno de estos sistemas es indispensable. En todo caso, es recomendable para el estudio del lenguaje en el niño.

La segmentación: el segundo paso

Una gran parte de los índices utilizados en el análisis del lenguaje (véase, por ejemplo, Rondal, 1980, 1983; Miller, 1981) implican la noción de frase y de enunciado. Un enunciado —aunque más tarde veremos que esta definición es insuficiente— ha sido a menudo definido como toda secuencia de sonidos comprendida entre dos pausas o interrupciones perceptibles en un flujo verbal. Una frase es un enunciado que comprende como mínimo un nombre o

pronombre y un verbo conjugado en concordancia gramatical con el sujeto. Las frases imperativas constituyen una excepción a esta regla, puesto que no son portadoras de sujeto, lo que permite distinguirlas de las frases construidas en indicativo. Es indispensable segmentar el discurso transcrito antes de cualquier análisis. La segmentación que nos interesa en este caso es la de los enunciados.

La informatización: el último paso

La entrada de un corpus de lenguaje en un ordenador no presenta dificultades especiales. Posee la inmensa ventaja de permitir el tratamiento rápido y eficiente de un gran número de datos recogidos y de cálculos que pueden efectuarse a partir de un corpus de lenguaje. Veamos en detalle el modo de realización de estas tres operaciones.

PRINCIPIOS DE TRANSCRIPCIÓN

En la investigación descrita, la transcripción estaba enfocada sobre las interacciones verbales entre una madre y su hijo, en un contexto de juego y conversación libre en casa. A la madre (M) y el niño (N) se unía a veces una tercera persona (PE: persona externa, fuera del binomio madre-hijo). Podía tratarse del padre, de la abuela materna o de una amiga de la familia.

Transcripción del discurso materno

Tanto desde el punto de vista de la sintaxis como del vocabulario, el discurso de un adulto como la madre se le considera correctamente desarrollado y organizado. No plantea ningún problema particular en cuanto a su transcripción. Para dar un reflejo tan fiel como sea posible en el plano fonético, se puede optar por el sistema AFI o por uno de sus derivados (véase más adelante). También se puede transcribir el discurso de la madre simplemente recurriendo al lenguaje escrito habitual, solución que hemos adoptado en nuestro caso. Sin embargo, para dar cuenta lo más fielmente posible de este lenguaje materno, hemos puesto de relieve:

1. *Las frecuentes apócopes.*
Ejemplos:
«Tu viens de l'prendre» por «Tu viens de le prendre».
«T'he dicho que no» por «Te he dicho que no».
2. *Algunas formas propias del lenguaje hablado.*
Ejemplos:

«Où est-c'qu'il est» por «Où est-il?».
«Qu'est-c'qu'on fait?» por «Que fait-on?».

3. *Palabras y expresiones regionales.*
Ejemplos:
«Mon poyon» por «Mon poussin». «Tirer son pull» por «Ôter son pull». «Tu n'saurais pas l'faire» por «Tu n'pourrais pas le faire».

4. *Las interjecciones y otras «rutinas» verbales.*
Ejemplos:
«bum», «aah», «pam», «pluf», «guau», etcétera.

Sólo hemos destacado estas producciones verbales cuando constituían una respuesta o un elemento de respuesta al interlocutor, y cuando su ausencia habría modificado el significado del anunciado. No existe ninguna codificación de las interjecciones (con la excepción de algunas «tradicionales» recogidas por los diccionarios). Su forma, número y significado no tienen otro límite que el de la imaginación del hablante. Con el fin de conservar una cierta coherencia en la transcripción, nos atrevernos a sugerir el código, implícitamente reconocido, de los comics o, en su defecto, recurrir al alfabeto fonético (o al sistema modificado descrito más abajo).

Transcripción del discurso de las PE

Es posible que personas distintas a la madre y al hijo intervengan en las grabaciones. Sólo hemos transcrito estas intervenciones en los siguientes casos:

1. PE- - →**N.** La PE se dirige especialmente al niño y obtiene de éste una respuesta.

2. PE- – →**M.** La PE se dirige a la madre. Estas intervenciones sólo son tomadas en cuenta si alteran el discurso del niño. Las reglas de transcripción son las mismas que rigen para el discurso de la madre.

Transcripción del discurso del niño

En la pronunciación del niño se dan ciertas imperfecciones. Para poder destacarlas fielmente, hemos decidido crear nuestro propio alfabeto fonético. Aunque esta nueva notación sufre el inconveniente de tener que representar ciertos sonidos con dos letras (y con dos signos en las tarjetas perforadas), posee, en relación al AFI, la doble ventaja de adaptarse a un material dactilográfico tradicional y presentarse visualmente más legible para el principiante.

ALFABETO FONÉTICO MODIFICADO (AFM)

i (perd*iz*)	p (*p*an)
e (p*e*lo)	t (*t*iempo)
a (g*a*to)	f (*f*oto)
u (l*u*to)	s (*s*in)
o (m*o*do)	k (*k*uando)
	l (*l*ento)
	r (*t*oro)
	b (*b*anco)
	d (*d*olor)
	v (*v*iento)
	z (*z*ona)
	g (*g*uante)
	ll (*ll*ano)
	rr (*r*ápido)
	m (*m*ina)
	ch (*ch*iste)
	in(g) park*in(g)*
	n (*n*ota)
	j (o*j*o, arpe*g*io)

En el alfabeto francés guardamos de la y (como en «yeux») únicamente su valor de yod. Así, se transcribe una palabra como «cocognes» (en la madre) como «kokony» (en el caso del niño).[1]

Ejemplo de transcripción del lenguaje del niño:

ma	**vè**	**zouwé**	avèk	mó	gá	tí
(moi)	**(vais)**	**(jouer)**	(avec)	(mon)	(grand)	(train)

1. Un ejemplo similar en castellano podría ser «pequeña/pekenya» *(N. del T.)*.

TRANSCRIPCIÓN DE LAS INTERJECCIONES Y OTRAS «RUTINAS» VERBALES

Con la excepción del modo de utilización del alfabeto fonético, los criterios son idénticos a los manejados en la transcripción del discurso materno.

Otras reglas de transcripción
(sin distinción de los interlocutores)

PROLONGACIÓN DE SONIDOS

Un punto (.) después del sonido significa la prolongación del mismo.

Ejemplo: «¿Poké.?» (N)

ACENTUACIÓN

Una palabra o una parte de la palabra fuertemente acentuada será subrayada.

Ejemplo:
M: «¡Cuidado!»
N: «Va zé un azidente.,

INTERFERENCIAS

Cuando se produce una interferencia entre ambos interlocutores, se señalará con una cruz (×) ahí donde se produzca.

Ejemplo:

M: «'s un barco muy **grande**» N: « gande»

PASAJES INCOMPRENSIBLES
(palabras, enunciados o parte/s de un enunciado)

Son registrados como (...) El símbolo (...) corresponde o a una palabra, si se ha podido calcular el número de palabras contenidas en el enunciado, o a una parte incomprensible de un enunciado.

Ejemplo: M: «(...) ya comió// (...) Ajá, ¿y me das todo, eso con tu (...)?».

LAS PAUSAS

Las pausas largas son registradas según su duración, con una o varias barras (/, //).

Ejemplo: M: «Ah, ¿esa es tu casa/ / ¡Qué bien!/ ¿Has ido muy lejos?»

ADEMÁS, hemos decidido considerar las secuencias demostrativas del tipo «celui-ci», «celui-là»,

como constituyentes de una sola palabra.[2] De forma paralela, las secuencias del tipo «voiture-là» han sido también registradas como una sola palabra y transcritas como «voiture-là» (es decir, con un guión entre ambos elementos).[3] Las palabras compuestas han sido transcritas como si se tratara de una sola, pero únicamente en el caso del niño. Por ejemplo, «sèwolá» (= cerf-volant),[4] «peutèt» (= peut-être), etcétera. Por el contrario, las secuencias del tipo «sabes tú»[5] serán contabilizadas como compuestas por dos o más palabras, lo que corresponde realmente a su composición en el nivel sintáctico. Estas decisiones se deben a que pudimos programar un ordenador para que se encargase del recuento de las palabras. En rigor, si el ordenador distingue, en una frase afirmativa como a(tú) lo dices», un pronombre (lo) y un verbo (dices), debe poder distinguir también estos mismos dos elementos en el giro imperativo «di (-) lo».

Ahora bien, la definición de *palabra* valida para el ordenador es la de «secuencia de caracteres precedidos y seguidos de un espacio en blanco». En el momento de informatizar nuestro corpus mediante tarjetas perforadas, hemos debido dar un tratamiento especial a aquellos espacios en blanco, cosa que era posible únicamente prescindiendo de las convenciones de la escritura (como, por ejemplo, el guión que suprimimos o el apóstrofo que separamos del resto mediante un espacio en blanco.

Ejemplo: Transcribimos:

«Est-ce ø que → «Est ø ce ø que
c'est» → c' ø est»

PRINCIPIOS DE SEGMENTACIÓN

Una vez terminada la transcripción de las grabaciones, procedemos a la segmentación del discurso en enunciados. Éstos estarán delimitados por sendas barras. Ejemplo: M: «/¿Vienes?/». Es difícil identificar y delimitar los enunciados en el flujo del discurso. Hemos definido el enunciado como «toda secuencia de sonidos comprendida entre dos pausas o bien interrupciones perceptibles en un flujo verbal».

Sin embargo, esta definición es demasiado teórica y general como para ser útil. Es necesario modificarla y precisarla.

En realidad, a menos de disponer de un material técnico muy perfeccionado, capaz de identificar y de medir con precisión las pausas en el discurso, es extremadamente difícil, cuando no imposible, apreciar debidamente la duración de las pausas que separan los elementos del discurso, salvo si estas pausas no están de por sí muy marcadas (1 segundo o más). También sabemos que la estructura sintáctica de los enunciados no deja de incidir en la percepción del discurso. Hemos acordado privilegiar el criterio gramatical a la hora de segmentar el discurso en enunciados. Concretamente, conservamos el criterio de pausa en la definición del enunciado, aunque añadimos algunos criterios gramaticales complementarios.

De esta manera, llegamos a una nueva definición fundada en cuatro criterios.

Definición del enunciado

CRITERIO 1

Un enunciado es una producción verbal marcada al comienzo y al final por una pausa claramente perceptible (distinta de una breve interrupción en el flujo verbal, o de una breve vacilación) o por una entonación particular claramente perceptible. Por ejemplo, las interrogaciones del tipo «¿Vienes?».

CRITERIO 2A

Si una unidad gramatical completa esta incrustada en una producción verbal más larga, sin la existencia de pausa, será considerada como enunciado separado del enunciado en que se encuentra incrustado.

Ejemplo: M: «Ah, eso / yo no estaba ahí / no lo sé».

2. El ejemplo no es traducible al castellano, puesto que las secuencias demostrativas de las que se habla para el francés no existen en castellano *(N. del T.)*.
3. Es el mismo problema *(N. del T.)*.
4. Una vez más, los ejemplos son intraducibles *(N. del T.)*.
5. En francés con un guión entre sujeto y verbo *(N. del T.)*.

CRITERIO 2B

Son considerados igualmente como enunciados distintos, las unidades gramaticalmente compuestas que se suceden sin estar separadas por pausas.

Ejemplo: «Hay que buscar un poco todas las piezas / hay que encontrar todas las piezas rojas / todos los ladrillos rojos como los del garaje/».

Consideramos igualmente como unidades gramaticales completas:

1: las frases, incluidos los verbos en imperativo;
2: las secuencias en las que el sintagma que contiene el sujeto gramatical ha sido omitido.[6]

Ejemplo:
N: «Quieo 'umo de na.nja mamá».
M: «Anda con cuidado, hijo».

CRITERIO 2C

Por razones puramente gramaticales, las proposiciones coordinadas y subordinadas son consideradas como constituyentes de un solo enunciado, a menos que estén separadas por una pausa o marcadas por un cambio en la entonación claramente perceptible.

Ejemplo: M: «/Podemos llevar muchos libros a la biblioteca / y entonces ¿qué hacemos?/».

Segmentación del discurso directo e indirecto

En este caso, la segmentación se hará entre el final del discurso indirecto y el comienzo del directo.

Ejemplo: M: «/ Me dijo / hay que.. /».

Repeticiones

Ahí donde hay repeticiones de elementos verbales que intervienen en el interior de una parte de la palabra, de una palabra o de un sintagma, contamos un solo enunciado.

Ahí donde las repeticiones envuelven unidades cuya extensión supera el sintagma, segmentamos en varios enunciados cuando se trata de repeticiones que tienen un valor funcional preciso, es decir, que no sean únicamente producto de una vacilación.

Ejemplo:
N: «/Poké, poké.../».
M: «/más clavos, más clavos... /».

En el discurso del niño, algunas repeticiones son claramente producto de la vacilación y de un dominio todavía imperfecto de la lengua, o de un control insuficiente del flujo verbal. En ese caso, contamos un solo enunciado.

Ejemplo: N: «Yo quieo ugá quieo ugá yo quieo ugá».

«Sí», «no» y otras rutinas verbales

Estas producciones son consideradas como enunciados sólo si constituyen una respuesta a una pregunta del interlocutor. El «mmhm», el «no» y el «sí», como repetición semántica, al igual que el «sí» como interjección no poseen, por lo tanto, valor de enunciados diferenciados.

Ejemplos:
M: «/¿La mamá ha subido?/»
N: «/Sí/»
M: «/¿Se va a su casa, sí?/».

Secuencias

Las secuencias de nombres o de verbos coordinados o no con continuidad semántica y forma interrogativa son considerados respectivamente como enunciados diferenciados.

Ejemplo: M: «/¿Es un fusil o un revólver?/¿un fusil?/¿grande o pequeño?/»

Palabras consideradas como enunciados

Es posible que en un diálogo una sola palabra sea considerada como enunciado. Esto sucede cuando:

— la palabra prolonga elípticamente el enunciado precedente del interlocutor.

Ejemplo:
M: «/Son grandes/ son/»
N: «gande».

6. Corriente en castellano, pero agramatical en francés *(N. del T.).*

— la palabra es la continuación de un enunciado precedente del hablante interrumpido por su interlocutor.

Ejemplo:
M: «Deja de moverte todo el tiempo alrededor de/»
N: «/¿del?/ »
M: «/de la mesa/»

— la palabra es una respuesta a una pregunta del interlocutor.

— la palabra es una corrección del hablante de lo ya dicho por el interlocutor, o una autocorrección.

Ejemplo:
M: «/¿Qué es eso?/»
N: «/Una kokony/»
N: «/Un pájaro/»
N: «/Un pajao/».

TABLA I. — *Extracto de un corpus de lenguaje mantenido por una madre y su niño en situación de conversación/juego libre.*

M_{134}/montre un peu ton bus chéri/regarde/ /(...) l'autobus qui roule sur le toit maint'nant/
E_{137}/í *twa*?/mè pas ke y a fè un aksida./
M_{135}/et tous les gens qui sont dans l'autobus sont blessés alors?/
E_{138}/mè nó i i só soti déza./
M_{136}/y avait personne dans l'autobus?/
E_{139}/mè i só déza soti pas ke y ôtobus s è ayèté/
M_{137}/m./oui mais le conducteur de l'autobus?/
E_{140}/i yè soti ôsi/
M_{138}/ben alors comment est ce que l'autobus a t il pu faire faire un accident?/
E_{141}/mè a vwatu k a fè ú aksidá avèk ôtobus/
M_{139}/l'autobus était arrête alors?/
E_{142}/ye ya dépeneus èl è fnu. è pi èl èl èn a épayé/ egat./
M_{140}/elle l'a réparé?/
E_{143}/egat/vè pát ma dépaneus//(...)
M_{141}/m./
E_{144}/egat ki sa dépaneus va fèr (...)/ó va apen apané* y ôtobus/
M_{142}/*dé*panner*/
E_{145}/*dé*pané/kom sa dépané/(...)//ó peu mèt tou sa ya su.?/
PE_5/parcee que/
E_{146}/pas ke kwa.?/
PE_6/parce que ça va ça va mieus/

Onomatopeyas

En términos generales, las onomatopeyas no son consideradas como enunciados diferenciados, excepción hecha de aquellos casos en que son claramente una respuesta dentro de un dialogo.

Ejemplo:
M: «/¿Cuántas ovejas hay?/»
N: «/Veja haze mm/ w
M: «/Beee/».

La *tabla I* ilustra, mediante la transcripción de un fragmento del corpus de lenguaje, el procedimiento empleado en una situación de conversación y juego libre entre una madre y su hijo.

La *tabla I* también explica la disposición espacial adoptada para la transcripción de los enunciados de la madre y del niño.

La distancia con respecto al margen en uno de los dos interlocutores permite una señalización más fácil de las respectivas producciones de ambos interlocutores (para el recuento de las intervenciones en la conversación, por ejemplo).

Los números que figuran delante de los enunciados del niño y de la madre permiten una señalización e identificación más fácil de estos últimos en el análisis.

INFORMATIZACIÓN DE LOS ENUNCIADOS

Resulta conveniente pensar en la informatización del corpus, especialmente si sus dimensiones son considerables. Desde luego, el ordenador puede realizar un cierto número de operaciones analíticas descriptivas, si se dispone de los programas o si es posible producirlos.

Por ejemplo, es relativamente simple programar un ordenador para establecer una muestra detallada de los índices de extensión media de producción verbal (EMPV —número de palabras por enunciado— véase Rondal, 1983) y ver su incidencia en diferentes fragmentos del corpus de lenguaje. En este caso preciso, resulta interesante obtener una muestra exhaustiva de las EMPV del niño y de la

madre por cada sesión grabada, durante un periodo de tiempo determinado, para el conjunto del corpus, etcétera, para poder establecer, por ejemplo, la evolución de la EMPV del niño, la estabilidad de ambos EMPV en función del número de enunciados tomados como base de cálculo, la evolución del primero en función de la edad, la evolución del EMPV de la madre en función de la del niño, etc.

Esto no es más que un ejemplo. Pueden utilizarse otros índices analíticos del lenguaje y una parte de ellos puede calcularla el ordenador (véase más adelante) bajo la condición, desde luego, de que el corpus haya sido anteriormente informatizado.

En nuestro trabajo hemos puesto en practica una serie de principios para la confección (perforación) de las tarjetas. Estos principios son validos únicamente si decidimos integrar los datos en el ordenador mediante la consola de su terminal.

Aun cuando en términos absolutos esta solución sea la más adaptable y racional, a veces no es la más indicada para las investigaciones largas y de gran envergadura, dado su elevado coste.

Tarjeta perforada

Contiene 80 posiciones disponibles para guardar la codificación de las informaciones destinadas al ordenador. Las seis primeras posiciones contienen, por este orden, las indicaciones relativas a: *el número de identificación de la grabación* (2 posiciones); *el número de la tarjeta perforada* (una por grabación, 3 posiciones); *la identificación del interlocutor* (Código: 1: Madre; 2: niño; 3: persona externa).

Aparece un solo interlocutor por tarjeta, lo que significa tener que pasar a la tarjeta siguiente (haya sido, o no, completada la precedente), cada vez que se produce un cambio de interlocutor.

Varias tarjetas sucesivas pueden ser atribuidas al mismo interlocutor, dependiendo de la extensión de su intervención (tarjetas en serie). Las posiciones 8 a 80 han sido reservadas para la transcripción del *corpus*, enunciado tras enunciado. Cada enunciado está situado entre dos barras oblicuas que permiten su recuento y suma (número de enunciados = número de barras menos uno) después de individualizar cada uno de los enunciados que pertenecen a una misma sesión de conversación.

Separación de las palabras

Para permitir la identificación de las palabras (y su recuento) mediante el ordenador, es indispensable que los sonidos constituyentes de una palabra se sucedan sin interrupción y que cada palabra se encuentre separada de la siguiente por un espacio en bIanco (una posición vacía). Cada palabra debe estar flanqueada por dos espacios en blanco, o precedida o seguida de una barra oblicua

Reglas relativas a las últimas posiciones de la tarjeta

Cuando una palabra debe segmentarse al final de la tarjeta, por falta de espacio, se continúa en la tarjeta siguiente, partiendo de la posición 7, sin barra oblicua ni guión. Cuando la última letra de la última palabra de un enunciado corresponde a la última posición en la tarjeta, la barra oblicua (o *slash*) con la que se debe terminar cada enunciado, se desplaza a la posición 7 de la tarjeta siguiente. Cuando una palabra de fin de enunciado termina en una de las tres penúltimas posiciones (77, 78, 79), va seguida de una barra oblicua y se continúa en la próxima tarjeta, aun cuando vaya seguida de otro enunciado.

De esta manera, no es necesario recortar la primera palabra de este nuevo enunciado. Las barras oblicuas que separan los enunciados no deben ir precedidas de espacios en blanco. Estos últimos se sitúan directamente después de las palabras o de los números de referencia que encontramos al principio de la tarjeta. Las palabras no pueden estar separadas por más de tres espacios en blaneo. Por último, no se pueden dejar más de tres posiciones después de la última palabra perforada al final de la tarjeta.

EL ALFABETO FONÉTICO EN EL ORDENADOR

Una de las exigencias que afecta por igual a la perforación y a la entrada de los datos en la consola del terminal es que el mensaje debe ser completamente «linealizado».

Puesto que el ordenador no permite la superpo-

sición de signos (por ejemplo, los acentos), debimos adaptar una parte de nuestro alfabeto fonético a esta limitación.

Procedimos entonces a las siguientes sustituciones:

í pasó a i ↄ
á pasó a a ↄ
â pasó a a ″
ú pasó a u ↄ
ó pasó a o ↄ
ô pasó a o ″
è pasó a e ⊂
é pasó a e ↄ

Esta opción, que depende de las posibilidades brindadas por el teclado de las perforadoras, no es, desde luego, la ideal. El codificador con poca experiencia se quejará en un primer momento de la aparición del signo ↄ para indicar a la vez nasalización y acentuación. Este inconveniente, que desaparecerá con la práctica, es inevitable, puesto que sólo puede utilizarse un número limitado de signos como símbolos fonéticos, ya que todos los demás signos tienen en el lenguaje informático un significado muy preciso.

OTRAS REGLAS

a) El signo de interrogación «?» debe ser perforado sin que medie un espacio en blanco entre la palabra precedente y el propio signo.

b) De la misma manera, el punto que señala la prolongación de un sonido debe ser colocado inmediatamente después del sonido en cuestión.

c) Las interferencias en las producciones verbales de los dos hablantes en interacción están señalados con el signo #, inscrito inmediatamente después de la palabra o de la parte del enunciado de cada hablante en el momento de la interferencia.

d) Una acentuación marcada es señalada con los signos < > a cada lado de la parte de la palabra, de la palabra o grupo de palabras acentuadas. *Ejemplo:* M: «/Está muy < bien >/».

e) Las interjecciones «hmm», «hum», etc., que indican la interrogación, la aprobación o el acuerdo en respuesta a una producción del hablante, son perforadas bajo la forma /m/, /m./ o /mm/.

f) Las secciones de una palabra, las palabras o partes de enunciados no entendidos están señalados por ... (entre dos barras oblicuas si se trata de todo un enunciado).

g) Una tarjeta formato precede a cada sesión grabada que forma parte del corpus. Se distingue de las tarjetas que le siguen mediante un asterisco * perforado en la primera posición, seguida de indicaciones que permiten identificar la sesión de grabación (número de la casete: posición 2 y 3; fecha de grabación: posiciones 4 a 10; número correlativo de la sesión de grabación: posición 12).

A estas alturas podrá surgir la pregunta de por qué creímos necesario adoptar un sistema de codificación (AFM) para el corpus de lenguaje, diferente del sistema finalmente empleado en el momento de la informatización del mismo corpus. Podríamos haber optado, efectivamente, por la transcripción directa del lenguaje grabado utilizando el mismo código que sirvió para la informatización.

Esta variante queda abierta para los que en el futuro recurran al sistema que nosotros hemos elaborado. La razón que nos lleva a preferir un trabajo en dos etapas, permitiéndonos tener acceso a dos sistemas de codificación, es la siguiente: en el caso en que la informatización del corpus de lenguaje no sea una etapa necesaria, se puede querer utilizar un sistema de transcripción lo más ajustado posible al lenguaje oral corriente o al alfabeto fonético en uso, conservando al mismo tiempo la posibilidad de mecanografiarlo mediante una máquina estándar. El AFM responde a esta doble característica.

INFORMATIZACIÓN DE LA TRANSCRIPCIÓN

Esta transcripción fue hecha gracias a un programa original capaz de tratar los datos en imágenes-tarjetas. Se trata de un programa escrito en VS FORTRAN. La opción por el lenguaje FORTRAN podría parecer sorprendente, dadas sus conocidas dificultades en el tratamiento de cadenas de caracteres. La opción fue definida tanto sobre la base del carácter portátil de este lenguaje como de las posi-

bilidades que nos ofrece. Hay que agregar que este programa funciona igualmente con los compiladores FORTRAN G y FORTRAN H.

Durante su ejecución, el programa identifica al hablante y yuxtapone continuamente, en un solo vector, el conjunto de tarjetas correspondientes a su discurso. Acto seguido, determina los diferentes enunciados y los imprime en secuencia (un enunciado por línea) de acuerdo con el identificador del hablante.

Si la extensión de un enunciado es superior al espacio disponible en el *listing* (80 caracteres, si consideramos que hay un espacio destinado al identificador del hablante) el programa determina el límite de impresión sobre el vector, y luego busca el primer espacio en blanco anterior a este límite. De esta manera, la impresión se encuentra siempre segmentada después de una palabra entera.

Los datos son anteriormente sometidos a un programa de validación que verifica la corrección de la estructura de las grabaciones y que facilita, si es necesario,. un cierto número de mensajes de error.

RECOLECCIÓN Y TRATAMIENTO INFORMÁTICO DE ÍNDICES

Aunque el punto que aquí trataremos no se inscribe en el marco de nuestra reflexión sobre la metodología de la transcripción, segmentación e informatización del corpus, nos ha parecido interesante abordar brevemente los problemas planteados por el tratamiento informático de ciertos índices. Aunque es verdad que el recurrir al ordenador y a su potencia de calculo ofrece a los investigadores unas ventajas innegables al ahorrarles, sobre todo, la larga y tediosa labor de recogida de datos, no se debe perder de vista el hecho de que el ordenador no puede sustituir en todos los casos al analista, dada su incapacidad de llevar a cabo una reflexión propia sobre el lenguaje.

En el caso que tratamos, el investigador no puede pedir asistencia de la máquina más que cuando se trata de operaciones matemáticas de recuento de enunciados, de frases o de palabras. Éste es el punto que deseamos abordar e ilustrar. Estas operaciones de recuento son aplicables a numerosos índices. Entre los que nos interesan más directamente está el cálculo de los índices de extensión media de los enunciados, el de la frecuencia distributiva de una serie de clases y de subclases formales y el de algunos índices conversacionales. La ilustración que sigue se refiere particularmente a estas dos ultimas categorías.

Recogida de datos

Resulta útil poder disponer de datos estadísticos referidos a la frecuencia de ciertas palabras y expresiones en el lenguaje intercambiado por los interlocutores.

A simple título de ejemplo, proporcionamos aquí una lista de 30 índices escogidos de una serie más amplia (de ahí su numeración del 129 al 163). Algunos se refieren a clases de palabras, otros nos remiten a algunas categorías conversacionales.

Índice	*Categorías*
129	Artículos definidos
130	Artículos indefinidos
132	Preposiciones
133	Conjunciones de coordinación
134	Conjunciones de subordinación
136	Pronombres personales de 1.ª persona
137	Pronombres personales de 2.ª persona
138	Pronombres personales de 3.ª persona
139	Pronombres relativos
140	Adjetivos indefinidos
141	Pronombres indefinidos
142	Adjetivos interrogativos
143	Pronombres interrogativos
145	Adjetivos demostrativos
146	Pronombres demostrativos de referencia próxima
147	Pronombres demostrativos de referencia distante
148	Pronombres demostrativos no referenciales
150	Adjetivos posesivos de 1.ª persona
151	Adjetivos posesivos de 2.ª persona
152	Adjetivos posesivos de 3.ª persona
153	Adjetivos posesivos que expresan la singularidad del poseedor
154	Adjetivos posesivos que expresan la pluralidad de los poseedores

Índice	*Categorías*
155	Adjetivos posesivos que expresan singularidad de lo poseído.
156	Adjetivos posesivos que expresan pluralidad de lo poseído
158	Pronombres posesivos de 1.ª persona
159	Pronombres posesivos de 2.ª persona
160	Pronombres posesivos de 3.ª persona
161	Llamadas de atención *(cuidado, mira, escucha)*
162	Aprobaciones verbales
163	Desaprobaciones verbales

Los índices 153 y 154 pueden combinarse con los índices 155 o 156 y/o los índices 150, 151 y 152. Véanse Rondal (1980) y Rondal y Neves (1978), para una explicación de los índices 150 a 156 y 161 a 163.

El procedimiento utilizado para la informatización de estos índices comprende dos etapas. La primera consiste en una identificación y clasificación sistemática de las palabras relativas a una u otra categoría. La segunda es la constitución de la biblioteca, y constituye un aspecto específico del tratamiento informático de los índices.

Identificación y clasificación de las palabras

Puesto que el ordenador es incapaz de realizar la más mínima distinción entre, por ejemplo, una preposición y un adjetivo posesivo, es necesario efectuar un censo de las palabras que pertenecen a cada una de las treinta categorías contempladas. Aunque el vocabulario de la madre no presenta ningún problema, al corresponder como lengua hablada por un adulto a las categorías gramaticales y a los diccionarios tradicionales, no sucede lo mismo con el lenguaje del niño. Éste domina imperfectamente tanto el léxico como la sintaxis o la pronunciación. Nos vemos ante la obligación de llevar a cabo una verdadera tarea de lexicografía cuando buscamos en el corpus las palabras del niño necesarias para el establecimiento de la biblioteca. Como consecuencia de la pronunciación imperfecta del niño, una misma palabra puede haber sido transcrita en el AFM de diversas maneras. Al contrario, esto puede dar lugar a numerosos casos de homofonía (por ejemplo ma = mon / ma / moi),[7] lo que añade complejidad a la tarea.

Después de una lectura atenta del corpus, hemos realizado una identificación sistemática de las palabras simples y de las expresiones compuestas por varias palabras del niño o de la madre susceptibles de pertenecer a una o varias de las categorías antes citadas. Luego hemos clasificado aquellas palabras o expresiones por grupos alfabéticos (o ficheros alfabéticos) de extensión creciente (de 1 a 23 letras). Con el fin de facilitar el recuento con el ordenador hemos considerado como letras los espacios en blanco (transcritos con el signo ø) que separan las palabras de una expresión compuesta, así como algunos signos específicos del alfabeto utilizado para la perforación.

Ejemplo: — et ø ensuite = expresión que contiene 10 letras.

Cada palabra va seguida del número de índice correspondiente (en este caso, de 129 a 163) y de un símbolo (M / N /) el cual indica si se trata de una palabra pronunciada por la madre o por el niño. El orden alfabético utilizado toma en cuenta las particularidades del alfabeto fonético adaptado a la perforación.

Orden alfabético:

1. ø (en blanco)
2. ´ apóstrofo
3. ⊃ Nasalización y cierre de e
4. ⊂ Apertura de e
5. ´´ Cierre de o
6. a
7. b
8. c

etcétera.

Ejemplo de clasificación: grupo de palabras de tres letras

esa	145M
mía	158M
po.	132N[8]

7. Su pronunciación en francés es muy parecida *(N. del traductor).*

8. No se trata de una traducción literal, sino de ejemplos en castellano (*N. del T.).*

CONSTITUCIÓN DE LA BIBLIOTECA

La biblioteca de índices está constituida por tantos ficheros distintos (montones de tarjetas) como diferentes extensiones tengan las palabras, es decir, 23. Cada fichero debe ser independiente de los demás, y se compone de tantas tarjetas perforadas como el número de palabras de una misma extensión. Estas tarjetas se perforaron de la siguiente manera:

— Posición 1 y 2: apuntador [el apuntador es la «extensión de la/las expresión(es) siguiente(s) (por orden creciente)»].

Definiremos la función más adelante.

— De la posición 3 a la posición *n*: palabra.

— A partir de la posición *n* + 1: categorías.

Cada categoría se perfora en tres posiciones puesto que su numeración corresponde a la lista de índices presentados más arriba. El número de categorías es variable, puesto que una misma palabra puede pertenecer a diferentes categorías (Ejemplo: «e ⊃» pertenece a las categorías 129, 130, 132, 133, 136 y 138).

Sin embargo, puede suceder que, al no ser más que el comienzo de una expresión compuesta, la palabra no esté repertoriada. Su notación será entonces øøø. (Ejemplo: (cual, quien) quiera).[9]

Funcionamiento

Sólo cuando este trabajo haya terminado, el ordenador puede empezar a tratar los datos, funcionando de la manera siguiente:

En un primer momento, el ordenador identifica una palabra del corpus, calcula su extensión y la compara a la de las palabras de igual extensión en el fichero. Si la palabra figura efectivamente en el fichero, hay un incremento en la categoría correspondiente. En ese momento, el apuntador permite al ordenador detectar y diferenciar:

a) Las expresiones compuestas cuya primera palabra es idéntica a la que ha sido identificada en el corpus y que además existe como palabra específica única en una de las categorías analíticas.

Ejemplo:

y	apuntador ø7
yøluego	apuntador 1ø
yøentonces	apuntador øø (puesto que se trata de la última expresión de la serie de expresiones compuestas por «y»).

Un apuntdor puede remitirnos a una secuencia perteneciente al mismo fichero (expresiones de la misma extensión).

Ejemplo:

después	apuntador ø9
después de	apuntador ø9
después de que	apuntador øø.

b) Las expresiones compuestas cuya primera palabra es idéntica a la que se ha identificado en el corpus, pero que no figura como palabra específica única, sin pertenecer, por tanto, a una de las categorías analíticas.

Ejemplo:

En cuanto a las palabras específicas únicas que figuran en la biblioteca y que no son las primeras palabras de una expresión compuesta, el apuntador es, en rigor, øø, puesto que se trata de palabras únicas que no se refieren a ninguna otra.

Ejemplo: Esteban apuntador øø.

RESUMEN

El artículo suministra las principales reglas de un procedimiento aplicable para la transcripción, la segmentación en enunciados y la informatización de un corpus de lenguaje.

El método presentado es de una gran utilidad en las etapas preliminares de preparación del corpus para el análisis psicolingüístico. De esta forma, sobre un corpus de lenguaje así preparado, se podrá efectuar un gran número de mediciones.

El artículo propone algunos elementos de análisis del corpus en clases formales y en categorías conversacionales con el único fin, sin embargo, de ilustrar la aplicabilidad de las técnicas informáticas disponibles.

9. Véase nota 8 *(N. del T.).*

Recibido: abril de 1985.