APLICACIóN DEL SISTEMA **MORFO** A UNA MUESTRA DE LENGUAJE INFANTIL

Giuseppe Cappelli (*), Victoria Marrero (+) y María José Albalá (++)

(*) Istituto di Linguistica Computazionale. CNR. Pisa.
(+) Universidad Nacional de Educación a Distancia.
(++) Consejo Superior de Investigaciones Científicas.

1. INTRODUCCIóN

En los últimos años han sido utilizados frecuentemente los instrumentos informáticos en el estudio del lenguaje infantil, especialmente tras la llegada del proyecto **CHILDES**[1] (**CHI**ld **L**anguage **D**ata **E**xchange **S**ystem), el cual incluye el sistema de codificación **CHAT** (Codes for the **H**uman **A**nalysis of **T**ranscripts) y el conjunto de programas **CLAN** (Computerized **L**anguage **AN**alysis), para el análisis cualitativo y cuantitativo de las transcripciones codificadas en CHAT.

Las ventajas de este proyecto son varias: contar con la experiencia previa de otros grupos de trabajo y con soluciones comprobadas para problemas comunes, disponer de datos recogidos de modo homogéneo en gran cantidad de lenguas, y, especialmente, la posibilidad de analizarlos automáticamente. Sin embargo, perduran algunas dificultades debidas a la naturaleza no estándar de la lengua hablada, que se acentúan en la fase de adquisición y desarrollo del lenguaje.

En España, nuestro equipo está desarrollando una base de datos del habla infantil a partir de grabaciones realizadas periódicamente a siete niños de edades comprendidas entre 1.4 y 6.6 años, desde enero de 1991 hasta ahora.

Desde un principio se hizo patente la necesidad de un instrumento fiable y económico que permitiera analizar morfológica y léxicamente estos ficheros, en los cuales no sólo encontramos un elevado porcentaje de palabras divergentes de la norma, sino también distintas realizaciones para la misma unidad. Llevarlo a cabo ha sido el objetivo del sistema **MORFO**, desarrollado en el Istituto di Linguistica Computazionale, con el fin de proporcionar una **lematización** y una **categorización morfológica** semiautomáticas que incluya las formas no estándar, superando así una de las limitaciones más importantes de otros sistemas de análisis morfológico semiautomático.

---

[1] MacWhinney, B.: The CHILDES Project: Tools for Analyzing Talk. **Hillsdale, New Jersey and London, Lawrence Erlbaum Associates, Pub., 1991.**

2. UN ANALIZADOR LéXICO Y MORFOLóGICO SEMIAUTOMáTICO

El conjunto de programas **MORFO** se integra en el entorno CHILDES desde varios puntos de vista: ha sido diseñado para trabajar con ficheros transcritos en formato CHAT; algunos de sus componentes pertenecen al paquete de programas CLAN, a los que se han añadido otros de nueva creación pero construidos con una filosofía similar a la de aquéllos; por último, los resultados que ofrece pueden ser procesados de nuevo (para búsquedas booleanas, recuentos selectivos de frecuencias, etc.) por cualquiera de las restantes aplicaciones de CLAN.

Para llevar a cabo los análisis se parte de la transcripción de las grabaciones sobre la que se realiza una serie de "barridos" sucesivos con el objeto de crear los siguientes ficheros de análisis:

1) Un léxico en el que aparecen todas las unidades producidas por cada informante ordenadas alfabéticamente (ficheros con extensión **.LES**).

2) Un "diccionario" morfológico[2] del informante que incluye las palabras del léxico anterior asociadas a sus correspondientes categorías morfológicas[3], y, en el caso de palabras no estándar, a su modelo (ficheros **.DIZ**).

3) Un fichero que contiene una línea de análisis morfológico[4] asociada a la línea de transcripción (línea principal) bajo la cual se coloca (ficheros **.MOR**).

4) Un fichero donde encontramos, además de las líneas principal y morfológica, una línea de "errores" que hace corresponder cada forma no estándar con su modelo (ficheros **.ERR**) y, en su caso, el código de error correspondiente.

---

[2]En realidad se trata de un léxico categorizado morfológicamente, no de un diccionario en el sentido lexicográfico.

[3]Para identificar la categoría morfológica de cada unidad, el programa batch "MINI" acude a un diccionario morfológico de referencia ("dizionar.mst") donde aparecen todas las etiquetas posibles para cada forma (por ejemplo, bajo la entrada cosa encontraremos el análisis como sustantivo y además como verbo).

Las entradas actuales del diccionario han sido extraídas de los ficheros codificados en CHAT disponibles hasta el momento, a las cuales se pueden añadir todas las que vayan apareciendo en transcripciones posteriores.

Actualmente existe el sistema MORFSIN: "Analizzatore Morfosintattico della Lingua Spagnola", realizado en el Istituto di Linguistica Computazionale del CNR de Pisa, por A. Saba, D. Ratti, M.N. Catarsi, y G. Cappelli. Este sistema consta de un diccionario de lemas (o radicales) que recoge todos los que aparecen en el Frequency Dictionary of Spanish Words (A. Juilland y E. Chang-Rodríguez, The Hague, Mouton, 1964).

Tenemos la posibilidad de desarrollar, de una manera económica, este diccionario de lemas para obtener un repertorio de formas flexionadas con sus correspondientes etiquetas morfológicas. En una segunda fase, proyectamos añadir el léxico infantil que no esté recogido.

[4]También en este caso utilizamos el "dizionar.mst" que, como mencionamos en la nota anterior, puede proporcionar varias etiquetas para una palabra, por lo que será necesaria una labor posterior de selección adecuada.

Aunque cada uno de estos pasos puede darse de forma independiente, con el programa batch "VIA" se han unificado en una sola orden:

VIA nombre-de-fichero INF

nombre-de-fichero = denominación del archivo de transcripción.

INF = código CHAT de identificación del informante.

[FIG. 1]

2.1. ELABORACIóN DEL LÉXICO

El léxico del informante se extrae del fichero que contiene la transcripción utilizando un programa batch (denominado "LES") que se vale de uno de los instrumentos estadísticos de CLAN (FREQ), pero sin asociar ninguna frecuencia a las entradas.

El sistema permite una aplicación selectiva de "LES", que incluya o excluya del análisis unidades determinadas: onomatopeyas, interjecciones, palabras de juego, neologismos, formas no identificadas, balbuceo, o, en general, cualquier cadena que el investigador seleccione.

2.2. ELABORACIóN DEL DICCIONARIO MORFOLóGICO DEL INFORMANTE

Se pueden seguir dos procedimientos: si no se van a analizar las palabras desviadas de la norma, el programa **MINI** relaciona la lista de las formas producidas por el informante con las correspondientes entradas del "dizionar.mst" (cfr. nota 2). En este caso, en el fichero **.DIZ** obtenido sólo las palabras estándar reconocidas serán etiquetadas, excluyendo todos los errores fonológicos, palabras de juego, neologismos, etc.

El segundo procedimiento permite analizar también las formas anómalas indicando al programa anterior que utilice un fichero (**.TAR**), confeccionado por el investigador, que ha de contener la lista de equivalencias 'forma anómala = modelo'.

Fichero: **INF1-291.TAR**

(l)a	**la**
(par)a	**para**
a(l)	**al**
abesa	**cabeza**
ahola	**ahora**
ala	**ahora**
bajala	**bajarla**
ejos	**conejos**
moj	**mejor**
ola	**ahora**

Este fichero de equivalencias tiene varias finalidades:

- permitir el etiquetado de formas anormales, haciendo corresponder las distintas realizaciones (ahola, ala, ola) de cada unidad (ahora);

- evitar los "falsos estándar" (en el ejemplo anterior, mojó sería identificado como forma del verbo mojar);

- facilitar la diferenciación entre formas anormales homófonas ([a] puede corresponder a la, al, para, o a; por eso en la transcripción se incluyen entre paréntesis los sonidos que faltan en cada caso);

- constituye la base del procedimiento posterior para crear la línea de "error", con una importante ventaja: los errores recurrentes (muy frecuentes en el lenguaje infantil) sólo deben incluirse una vez en el fichero **.TAR**; posteriormente, y de modo automático, se insertarán en la línea de error cada vez que aparezcan.

El fichero **.DIZ**, obtenido con cualquiera de los dos procedimientos anteriores, presenta los datos del siguiente modo:

Fichero: **INF1-291.DIZ**

..	**...**
anda	**E111C E121B andar**
animales	**A1+ B3+ animal**
aquinas	**gallinas**
aquí	**F00L aquí**
asul	**azul**
así	**F00M así**
...	**...**

Las palabras identificadas por el diccionario (anda, animales, aquí, así) aparecen

con todas sus posibles categorías[5]; las formas anómalas (aquinas, asul) con su modelo.

De esta manera resulta posible unificar todas las palabras recogidas según los criterios lexicográficos habituales (todas las formas verbales reunidas bajo el infinitivo, las variaciones de género y número bajo el masculino singular, así como los diminutivos y las distintas derivaciones de una misma unidad bajo la entrada correspondiente).

2.3. ANáLISIS MORFOLóGICO

El programa **MORF** construye, a partir del fichero de transcripción y del diccionario morfológico asociado a éste (**.DIZ**), un nuevo archivo (**.MOR**) que añade la línea de etiquetado morfológico a la de transcripción:

Fichero: **INF1-291.MOR**

[...]
***INF: ah, bien, voy a dala a momé una fo; ay!**
%mor: <1> $EXCL ah
<2> ADVMDL bien $CONJ$MDL bien NMAS$SG bien
<3> $V3$IND$PRES$1S ir
<4> $PREP a
<5> $V1$INF dar + PROFEMSGPER$3S$A la
<6> $PREP a
<7> $V2$INF comer
<8> ARTFEM$SGb un PROFEMSGINDEF un ADJFEMSGNUM uno PROFEMSGNUM uno $V3$SUB$PRES$1S unir $V3$SUB$PRES$3S unir
<9> NFEM$SG flor
<10> $EXCL ay
***INV: vas a coger una flor?**
***INF: sí.**
%mor: <1> ADVAFF sí PROINVSGPER$3S$T sí
[...]

Entre ángulos aparece el "locus" (número de orden de la palabra en la frase) de cada forma para facilitar la correlación con la línea principal.

Una vez que, en las palabras ambiguas, haya sido escogida la opción adecuada a cada enunciado (en una, por ejemplo), y se hayan completado las etiquetas que faltan, el fichero estará listo para ser elaborado por los programas CLAN, que ofrecerán la información cuantitativa y cualitativa necesaria para conocer el desarrollo léxico y morfosintáctico de cada informante.

---

[5]Los códigos que acompañan a las palabras etiquetadas corresponden a cada categoría morfológica (por ejemplo,
E = verbo, **E111C** = verbo de la 1ª conjugación, indicativo, presente, 3ª persona del singular).

2.4. CREACIóN DE LA LíNEA DE ERROR

El programa **CREAR** permite insertar de modo automático una línea secundaria, además de la morfológica, donde, como hemos dicho, aparecen las formas desviadas de la norma con sus modelos correspondientes. **CREAR** toma como ficheros de entrada tanto el de transcripción como el de equivalencias.

Fichero: **INF1-291.ERR**

[...]
***INF: ah, bien, voy a dala a momé una fo; ay!**
%err: <5> dala = darla
<7> momé = comer
<9> fo = flor
%mor: <1> $EXCL ah
<2> ADVMDL bien $CONJ$MDL bien NMAS$SG bien
<3> $V3$IND$PRES$1S ir
<4> $PREP a
<5> $V1$INF dar + PROFEMSGPER$3S$A la
<6> $PREP a
<7> $V2$INF comer
<8> ARTFEM$SGb un PROFEMSGINDEF uno ADJFEMSGNUM uno PROFEMSGNUM uno $V3$SUB$PRES$1S unir $V3$SUB$PRES$3S unir
<9> NFEM$SG flor
<10> $EXCL ay
[...]

Este programa es adicional, y su utilización es facultativa e independiente del procedimiento para obtener el análisis morfológico. Resulta especialmente útil para el estudio de la adquisición y desarrollo de la fonología.

3. APLICACIONES EN EL ESTUDIO DEL LENGUAJE INFANTIL: ANáLISIS DE FORMAS DIVERGENTES DE LA NORMA.

Hemos aplicado los programas **MORFO** a doce de nuestros ficheros, con la finalidad de establecer la eficacia del procedimiento para analizar las formas no estándar. El corpus consta de un total de 15.057 palabras, distribuidas de la siguiente manera:

Infor-mante	Edad (*)	Sexo	Fecha grabación	nº palabras / grabación	Total palabras
INF. 1	2,2	MASC	feb. 91	962	
	2,5		may. 91	1772	
	2,9		sep. 91	858	3592
INF. 2	2,8	FEM.	sep. 90	340	
	3,0		ene. 91	748	
	3,4		abr. 91	827	
	3,8		ago. 91	1428	
	4,0		ene. 92	909	
	4,8		ago. 92	1333	5585
INF. 3	3,5	MASC	abr. 91	2639	
	3,7		jun. 91	2175	
	4,0		nov. 91	1019	5833

(*) Expresada en "años, meses" en el momento de la grabación.

Todos los ficheros (en formato CHAT) fueron procesados, en primer lugar, utilizando el "dizionar.mst" (cfr. nota 3), pero sin aplicar el procedimiento de identificación de formas no estándar (es decir, del modo en que lo haría cualquier sistema de análisis morfológico por ordenador). El porcentaje global de palabras analizadas así fue un **14,58%** menor que el que obtuvimos al incorporar el programa que asigna las correspondientes formas estándar a las realizaciones anormales del niño.

Consideramos, por consiguiente, que este último proceso consigue optimizar, casi en un **15%**, la eficacia de un analizador morfológico semiautomático aplicado al lenguaje infantil, y, por extensión, a cualquier realización lingüística desviada de la norma.

Naturalmente, las ventajas de un procedimiento para identificación y análisis de formas no estándar decrecen a medida que el habla analizada se aproxima a la norma lingüística: las cifras globales anteriores esconden marcadas desigualdades debidas a las diferencias en edad y nivel de desarrollo lingüístico de los distintos informantes.

INFORMANTE	EDAD	DIFERENCIA * %
INF1	2,2	11,02
	2,5	20,37
	2,9	10,72
INF2	2,8	33,82
	3,0	19,12
	3,4	11,73
	3,8	21,78
	4,0	20,79
	4,8	19,05
INF3	3,5	2,58
	3,7	2,34
	4,0	2,75

* Diferencia entre análisis con equivalencias y sin equivalencias

El informante 3 es un niño caracterizado por una notable madurez lingüística; esto explica la poca diferencia entre los dos tipos de análisis. Podemos considerarlo un prototipo de sujeto escasamente desviado de la norma: en él el analizador morfológico consigue los mayores porcentajes de palabras etiquetadas con cualquiera de los dos procedimientos.

Con los otros dos informantes ocurre lo contrario: en el primer caso un incipiente desarrollo del sistema lingüístico hace que la proporción de palabras no estándar sea muy elevada; en el segundo caso ocurre lo mismo, aunque por razones más complejas (diferente entorno dialectal, menor estimulación externa, etc.).

4. CONCLUSIONES

El sistema **MORFO**, que hemos presentado aquí, es un instrumento para realizar un análisis morfológico y léxico de manera rápida y económica.

A partir de un fichero de transcripción, el programa **LES** crea un léxico ordenado alfabéticamente de las formas utilizadas por cada informante. A continuación, el programa "MINI", consulta este archivo, además de un diccionario de referencia y, en caso de habla no estándar, un fichero de equivalencias entre las formas anómalas y sus modelos; esto le permite obtener una línea de análisis morfológico. Si interesa codificar y analizar las anomalías que presenta el informante, el programa **CREAR** proporciona una línea secundaria sobre la cual aparecen los errores y sus equivalencias, unidas, en su caso, al código que les corresponda.

MORFO no pretende agotar las posibilidades de un análisis morfológico como finalidad en sí, sino ser un instrumento útil integrado en un proyecto más amplio (CHILDES). Esto permite reelaborar los datos obtenidos para efectuar análisis cuantitativos y cualitativos con los programas del paquete de aplicaciones CLAN y aprovechar una base de datos interlingüística homogénea. Todo ello hace posible el intercambio y análisis morfológico contrastivo entre distintas lenguas.

Se trata, además, de una herramienta especialmente adecuada para el análisis de lenguajes desviados de la norma (como es el caso del lenguaje infantil), puesto que permite, entre otras ventajas, establecer una correspondencia entre distintas formas de la misma palabra, evitar los falsos estándar, diferenciar formas anormales homófonas y etiquetar automáticamente errores repetidos.