Análisis automático de la gramática temprana [1]

Automatic analysis of early grammar

Zulema G. Solana
GRUPO INFOSUR
UNR
zsolana@arnet.com.ar

Abstract

This work aims to provide a general view about the syntactic and morphological aspects of early grammar and to determine some parameters to implant it into the machine. First, the order in which the categories N (noun), A (adjective), V (verb) and P (preposition) appear in the child and then, the system of morphological and semantic inflectional marks of these categories are determined. The study of prefixes and suffixes which may appear in word formation and of a more specific determination of the evolution stages is left for a later stage in the analysis. Only one sample from three children between the specified ages has been collected. They belong to the CHILDES Data Bank and the resulting generalizations are of a provisional nature.

Keywords: early grammar, morphological inflectional marks, semantic marks

Resumen

Este trabajo se propone presentar una descripción general de la gramática temprana en sus aspectos morfológicos y sintácticos y determinar algunos parámetros para implantarla en máquina. Se comienza por establecer el orden en que aparecen en el niño las categorías N(nombre), A(adjetivo), V (verbo) y P (preposición) y el sistema de marcas morfológico-semánticas de flexión de estas categorías. Se deja para más adelante el estudio de la aparición de sufijos y prefijos en la formación de palabras y una determinación más específica de las etapas de evolución. Se ha tomado una muestra de sólo tres niños, en las edades consideradas, pertenecientes al Banco de datos de CHILDES, razón por la cual las generalizaciones tienen carácter provisorio.

Palabras clave: gramática temprana, marcas inflexionales morfológicas, marcas semánticas

1. INTRODUCCIÓN

Me propongo presentar una descripción general de la gramática temprana[2] en sus aspectos morfológicos y sintácticos y determinar algunos parámetros para implantarla en máquina. Comenzaré por establecer el orden en que aparecen en el niño las categorías N(nombre), A(adjetivo), V (verbo) y P (preposición) y el sistema de marcas morfológico-semánticas de flexión de estas categorías. Se deja para más adelante el estudio de la aparición de sufijos y prefijos en la

[1] Este trabajo pertenece al PID de Ciencia y Técnica 2012 del mismo nombre
[2] Con "gramática temprana" se hace referencia al conocimiento lingüístico de las primeras etapas del desarrollo

formación de palabras y una determinación más específica de las etapas de evolución. Se ha tomado una muestra de sólo tres niños, en las edades consideradas, pertenecientes al Banco de datos de CHILDES, razón por la cual las generalizaciones tienen carácter provisorio.

2. ESTADO ACTUAL DE LOS CONOCIMIENTOS SOBRE EL TEMA

En lo que respecta a las investigaciones sobre la informatización del análisis del lenguaje infantil hay que mencionar en primer lugar a todo lo producido en el marco de CHILDES. La base de datos CHILDES[1] (Child Language Data Exchange System) contiene 44 millones de palabras de 28 lenguas diferentes. La información contenida en el sistema corresponde a 4500 investigadores que trabajan sobre éste y realizan sus contribuciones.

Los datos corresponden a transcripciones de conversaciones de niños con sus madres, padres, investigadores, etc. Todos estos archivos de transcripciones, pertenecientes al sistema CHILDES, se encuentran en el formato CHAT (Codes fr de Human Analysis of Transcripts). El sistema CHAT provee un formato estandarizado para producir transcripciones informatizadas de conversaciones. Estas conversaciones involucran niños y padres, niños, profesionales/investigadores y padres o maestros.

Los objetivos de un sistema informático para datos de este tipo pueden resumirse en:

a) automatizar el proceso de análisis de los datos

b) obtener mejores datos en un sistema de transcripciones consistente y completamente documentado

c) obtener más datos de una mayor cantidad de niños de variadas edades y lenguas

Estos archivos con formato CHAT pueden ser analizados mediante el programa CLAN (Computerized Language ANalysis). El programa fue diseñado para facilitar las investigaciones y análisis sobre las transcripciones en formato CHAT.

En nuestra investigación acudiremos al banco de datos de CHILDES, pero analizaremos los datos con nuestras propias herramientas informáticas.

3. METODOLOGÍA Y ETAPAS

Se trabajará con lenguaje espontáneo dejando para una etapa posterior la elaboración de test o pruebas, es decir, se recurrirá por el momento a la observación dejando para después el uso de métodos experimentales.

La posibilidad de guardar en base de datos informatizadas grandes corpus de lenguaje infantil hace posible el trabajo con el lenguaje espontáneo, en general producto de conversaciones de adultos con el niño o de niños entre sí, metodología (la de la observación del lenguaje espontáneo) que en décadas pasadas había sido desplazada por la experimentación ya que, no mediada por grandes corpus informatizados, la observación era un proceso costoso en tiempo.

Para la implantación en máquina se recurrirá a las herramientas que tiene el grupo INFOSUR [2].

3.1. Herramientas informáticas con que se cuenta

- El software Smorph, analizador y generador, realiza la tokenización y el análisis morfológico, en una sola etapa y da como resultado las formas correspondientes a un lema con sus valores. Lo presentó en su tesis doctoral Salah Aït- Mokthar [3]. La tesis fue desarrollada en la Universidad Blaise-Pascal de Clermont-.Fd bajo la direción de Gabriel G. Bès.

Consideramos que el proyecto INFOSUR, desarrollado en el ámbito de nuestra universidad, es no sólo un antecedente sino la base que nos proporciona loa elementos de partida en el aspecto informático y el trabajo de modelización del español. [4]

- MPS trabaja SMORPH en una organización modular. La salida de smorph, en lenguaje Prolog, es la entrada de MPS, también procede de una tesis(cf. Abacci) dirigida por el Dr. Bès en Clermont-Fd.

- xfst (Xerox Finite State Tools): Esta herramienta informática ha sido usada por Xerox Research Centre Europe (XRCE) y Palo Alto Reserach Centre (PARC.) Es un autómata de estados finitos en el que se ingresan las propiedades lingüísticas en forma de reglas, que pueden ser testeadas en el proceso de generación.

3.2. Hipótesis de partida

Respecto del sistema de marcas morfológico-semánticas de las categorías N, V, A, ADV, P, PRON, se plantea la siguiente hipótesis:

> Respecto del conocimiento y uso de marcas morfológicas de flexión, los niños pasan por tres momentos:
>
> > -categorías sin marcas de flexión: por ejemplo, *nene* aplicado a masculino, femenino, singular y plural, o *come* aplicado a primera o tercera persona singular.
> >
> > -primeras marcas en algunas categorías: por ejemplo diferenciar persona con *como* y *come.*
> >
> > -un desarrollo morfológico complejo

3.3. Corpus

Para esta etapa de la investigación, el corpus consiste en una muestra [5] procedente de expresiones espontáneas de niños hablantes de español menores de cuatro años del Banco de Datos CHILDES.

La investigación tiene propósitos metodológicos y no va a evaluar precisión y cobertura de la implementación de las herramientas informáticas sino sólo la posibilidad de ser empleadas en adquisición del lenguaje.

3.4. Etapas

Consideraremos dos etapas:

3.4.1. Antes de los dos años

3.4.1.1. Categorías morfológicas

Predominan los nombres y las primeras producciones son palabras bisílabas, sin sufijos ni flexivos ni derivativos Ejs*: nene, pupa, tete, pie, guaguau, agua, peíto*(por *pelito*), *bota, calle*, *silla , ponja (*por *esponja).*

Posteriormente aparecen antecedidos en varios casos por determinantes[6]: Ejs: *e nene, a calle, a silla, *este silla, a botas.* Aquí estoy considerando como determinantes propios de esta etapa a (e/o/a/a) que explicitan los rasgos que no están gramaticalizados dentro de N.

Finalmente antes de los dos años aparecen nombres trisílabos. Ejs: *cabeza, mañana, chaqueta* y se

enriquecen los determinantes, con formas de los indefinidos un/una/unos/unas, propios de la gramática adulta, pero, en algunos casos con variaciones fonéticas: *u' botó' , un caballo*

En cuanto a los verbos, se reafirma lo que se sabe por la literatura especializada [7] se encuentran verbos aparentemente en tercera persona singular, pero en realidad sin persona asignada desde el momento en que no constituyen ninguna oposición. Ejs: *está, no´sta, pincha, chupa, gusta, come, quita* luego *hay, abe(*por *abre), ueve(* por *llueve), vio, caió (cayó), pasó*. Quedan para remarcar dos casos: un verbo irregular diptongado en primera y tercera singular: *siento, sienta* y *cae* acompañado de *se*, que es el primer clítico registrado *se cae* El hecho se reafirma un mes más tarde con *se acabó*

Al final de esta etapa aparece el clítico de primera persona *me,* se refuerza el *se* y ambos aparecen con verbo en pretérito perfecto. *m'a chupa'o, s' ha perdido, s' ha loto* .

Adjetivos solamente: *malo*, *bonito* y *caente* (por *caliente)*

3.4.1.2. La sintaxis antes de los 2 años

Vamos a caracterizar la sintaxis de la gramática temprana a partir de los siguientes ejs:

a.Se cae e nene

b.Papá a calle

c.E´nene a botas

d.Quita guauguau

e.Sienta mamá

f.Siento aquí

g.Echo agua

h. *Ota(otra) vez cayó*

i.No apabó

j. *Mía(mira) mama mi pie.*

k. No hay papú (champú)

l.Mamá, m'a chupa'o e' guauguau .

ll.S' ha perdido e' pendiente .

Si se toma en consideración sintagmas, sintagmas núcleos y oración pueden hacerse algunas observaciones de conjunto y otras particulares:

-La mayor parte de las oraciones cuentan entre dos y cuatro palabras.

-Algunos sintagmas están formados por N solo, otros por det+N

-Entre los determinantes, se encuentran posesivos(*mi* en j), indefinidos (*ota* en h), artículos (*e, a* en a,b y c)

-Hay SN sujetos y objetos directos.

-Aparición de verbo en primera persona(e), lo que habilita a decir que (f) está en tercera, dado que hay una oposición de persona con el mismo verbo. La primera (a) con un verbo de clítico obligatorio, otro clítico en l (me (m´) es el objeto de *"a chupa'o")*y en (ll) aparece la primera

expresión de impersonalidad.

3.4.2. A los 2 años

3.4.2.1 .Categorías morfológicas

A los 2;02 los Adjetivos. *secos***,** *mojado, mojados, buena, buenísima, guapa, fío (*por *frío), malo, bonito, pequeño, novo (por nuevo)*

Sufijos derivativos, sólo el diminutivo : *sapatito, bonito a los 2;02.*

Sistema verbal María 2;00- 2;03

dej -o v/pres/ind/1a/sg

-as v/pres/ind/2a/sg

-a v/pres/ind/3a/sg

dej -e v/pres/subj/1a/sg

-es v/pres/subj/2a/sg

-e v/pres/subj/3a/sg

Clíticos María 2 a 2;03

me	se me
te	te los
le	se le
lo	me lo
se	

3.4.2.2. La sintaxis a los 2 años

2;02

a.*N**o, no te gusta las galletas*** .

b.Papá, tú que me, ***tú me*** *que(d)as conmigo, a qué sí ?*

c.***Te*** *vo a mojar,* ***a Papá*** .

d.*Venió un bolo feroz .*

e.*Estaba una niña guapa .*

f.*Estaba una niña gua:pa que se llama Maniña y viene un bolo feroz*

Hay falta de concordancia en (a), en la persona del clítico (*me* en lugar de *te*)en (b), *"venió"* por "*vino*" en (d). Aparece una relativa y una coordinación en (f)

2;04

a.*Espera que voy a coger un muñeco .*

b.*Hacemos una película al Aito .*

c.*Si no jubamos la tiro la tapa .*

d.Yo sabo hasé muchas casas

e.*Ahora voy a hasé* ***ot'a casa más bonita***

Puede observarse que se va aumentando el número de palabras por oración, que hay subordinadas, entre ellas una condicional en (c). Crece la variedad de determinantes (*muchas casas, ota casa)* y continúa la conjugación regular donde en la lengua estándar es irregular (*sabo).* Llama particularmente la atención un SN como ***ot'a casa más bonita*** (indef + N + SADJ (adv + adj)). En esta etapa se observa la concordancia correspondiente a la lengua adulta.

3.4.3. Síntesis de la evolución

	Antes 2 años	2 a 2;6
morfología N	ausencia flexión y derivación	flexión plural concordancia género y número en el SN
Determinantes	artículos (*e/a/o/a*) indefinidos (*ot(r)o/a, u(n))* posesivos (*mi*)	artículos (*la, las*)
Adjetivos	*malo, bonito, ca(li)ente*	*mojado, mojados, buena, buenísima, guapa, fío (*por *frío), pequeño, novo(por nuevo)*
morfología V	3a-pers.sg.presMI : *está, no´sta, pincha, chupa, gusta, come, quita*, *hay, abe(*por *abre), ueve(* por *llueve),sienta.* 3a-prts.sg.presMI *vio, caió (cayó), pasó.* 1a-pers.sg:*siento* 3a-pers.sg.pret.perf.MI. *m'a chupa'o, s' ha perdido, s' ha loto*	1a-2a-3a.pers-sg-MS 1a.pers.fut.MI : voy a + infinitivo
Clíticos	*me/se*	*te/le/lo/se me/ se le/te lo/me lo*
compl verbo	OD	
Prep	-	*a, con, para*

4. IMPLANTACIÓN EN MÁQUINA

Implantación de la morfología del español estándar

El software Smorph, mencionado en 3.1., realiza en una sola etapa la tokenización y el análisis morfológico. A partir de un texto de entrada se obtiene un texto lematizado con las formas correspondientes a un lema (o a un subconjunto de lemas) con los valores correspondientes. Es una herramienta declarativa, la información utilizada por Smorph está separada de la maquinaria algorítmica, esto hace que se la pueda adaptar al uso que quiera darse, de modo tal que con el mismo software se puede tratar cualquier lengua siempre y cuando se modifique la información lingüística declarada en sus archivos.

Se ha trabajado con información pertinente para el francés, para el portugués y para el español [6] En todos estos casos se trata de la lengua estándar, en este trabajo daremos cuenta de la investigación que estamos realizando respecto de las producciones de la gramática temprana. A continuación explicaremos como se implementa en general y luego las modificaciones realizadas para que se puedan analizar las producciones infantiles.

Esta herramienta compila, minimiza y compacta la información lingüística de modo que quede disponible en un archivo binario. Los códigos fuente se dividen en cinco archivos:

- Códigos Ascii
- Rasgos
- Terminaciones
- Modelos
- Entradas

En el archivo **entradas**, se ingresan los ítemes léxicos acompañados por un indicador del modelo correspondiente. Este indicador de modelo oficia de enlace con el archivo **modelos,** donde se especifica la información morfológica, género y número y las terminaciones que se requieren en cada ítem.

En el archivo modelos, se introduce la información correspondiente a los modelos de flexiones morfológicas. Un modelo de flexión agrupa todas las flexiones de una misma clase de palabras. Esto se describe asociando a un conjunto de terminaciones el correspondiente conjunto de definiciones morfológicas. El esquema para definir los modelos es el siguiente:

<nombre_modelo> -<cantidad de caracteres a sustraer>

<terminación 1> <definición morfológica para terminación 1>

<terminación 2> <definición morfológica para terminación 2>

...

<terminación k> <definición morfológica para terminación k>

Se declara en primer lugar el nombre del modelo, luego se declara la cantidad de caracteres que hay que sustraer a la forma lematizada. Este valor debe ser una cifra entre 0 y 9 y estar precedida del signo "-". En tercer lugar se declara la terminación, que debe estar declarada previamente en el archivo de **terminaciones**. La declaración morfológica corresponde a una cadena de caracteres sin espacios en blanco.

En el archivo de **terminaciones** es necesario declarar todas las terminaciones que son necesarias

para definir los modelos de flexión. Si en la definición de un modelo se especifica una terminación no declarada en este archivo, el programa emite un mensaje de error. Las terminaciones se declaran una a continuación de otra, separadas por un punto. Es posible declarar una terminación vacía mediante el carácter "@" y una terminación distinguida asociando a una terminación la definición morfológica correspondiente.

Para construir los modelos se recurre a rasgos morfológico- sintácticos (categoría, género, número, etc). En el archivo de **rasgos**, se organizan jerárquicamente las etiquetas, por ejemplo, nombre, adjetivo, etc. Asimismo, se puede incorporar la etiqueta que indica, por ejemplo, el tipo de nombre y se adicionan los rasgos de concordancia, género y número:

En el archivo de códigos **ASCII** se especifican, entre otros, los caracteres separadores, las equivalencias entre mayúsculas y minúsculas.

El archivo **data**, contiene los nombres de cada uno de los archivos descriptos anteriormente.

A continuación se presentará la implementación de los archivos nombrados, para el análisis de ejemplos pertenecientes al español estándar y luego **los cambios necesarios para su adaptación de modo de poder realizar el análisis de expresiones de la gramática temprana.**

Hemos encontrado en la gramática de los 2 años el siguiente paradigma para los artículos:

e/a/o/a

En el español estándar tenemos:

Entradas

el el /det/art .

el los /det/art .

el la /det/art .

el las /det/art .

el lo /det/art .

En cada fila, primero el lema (elegimos convencionalmente “el”), luego la ocurrencia lingüística y después los rasgos (“det”: determinante, “art”: artículo).

Para poder analizar las ocurrencias lingüísticas infantiles, introducimos:

el e /dett/artt .

el o /dett/artt .

el a /dett/artt .

el a /dett/artt .

Equivale “dett”:determinante gramática temprana y “artt”:artículo gramática temprana.Además de modificar el rasgo en “entradas” tenemos que agregarlo en el archivo “rasgos”.

Así, si analizamos “e nene”(el nene), obtenemos:

'e'.

['el', 'EMS','dett', 'TDETT','artt'].

'nene'.

['nene', 'EMS','nom', 'GEN','masc', 'NUM','sg'].

-cuestiones fonológico- fonéticas que llevan a duplicar las “entradas”

Ejs: ota/otra, fío/frío, apabó/acabó, jubamos/jugamos

-cuestiones morfológicas que llevan a modificar los “modelos”

Ejs: sabo/sé, venió/vino

Aquí se trata de la conjugación de verbos irregulares como si fueran regulares.

“venir” como regular sigue el modelo 3 ej.”partir” (vení/veniste/venió)

@v3 -2

+í v/prets/ind/1a/sg/c3/r

+iste v/prets/ind/2a/sg/c3/r

+ió v/prets/ind/3a/sg/c3/r

(rasgos: V(verbo), prets(pretérito simple), ind(indicativo), 3ª(tercera), sg(singular), c3(3ª conjugación), r(regular))

Para que este modelo, propio de “partir/partió”, se adapte a venir/venió, debe convertirse en un nuevo modelo, al que asignamos otro número:

@v4 -2

+í vt/prets/ind/1a/sg/c3/r

+iste vt/prets/ind/2a/sg/c3/r

+ió vt/prets/ind/3a/sg/c3/r

Cambiamos el rasgo “v” por “vt” (verbo gramática temprana) y, en consecuencia en el archivo “rasgos” hay que agregar el nuevo rasgo.

A MODO DE CONCLUSIÓN

Respecto de la evolución sintetizada en 3.4.3 puede decirse que en el SN se parte de una forma invariable que pronto toma variaciones de género y número, lo que permitirá el proceso de concordancia. Cuando todavía N no flexiona ya presenta un determinante (e/a/o/a) que anuncia su género y número en una especie de morfología flexional pre-posicionada. Los adjetivos avanzan en un sentido cuantitativo y cualitativo, en la segunda etapa se triplican y aparecen los participios adjetivales.

El verbo también parte de una forma invariable hasta lograr primero oposición de persona y en segundo término de tiempo. Se establece el sistema de clíticos que incide en las posibilidades sintácticas de la lengua.

En la propuesta de implementación en máquina hemos duplicado entradas por razones fonológico-fonéticas y hemos modificado modelos para dar cuenta del tratamiento de verbos irregulares como si fueran regulares. Esta reestructuración del sistema lleva a agregar nuevos rasgos. El hecho de contar con una investigación e implantación en máquina de la morfología del español estándar nos permite la implementación presentada para la gramática temprana. Sólo cuando ampliemos la muestra estaremos en condiciones de evaluar la precisión y cobertura de lo realizado.

Referencias

[1] Carrasco González, M. y Celis Sánchez, C. (2004): CHILDES Project: Child Language Data Exchange System. Sistema de Transcripción CHAT. Publicación electrónica disponible en http://childes.psy.cmu.edu/intro/spanish.pdf

[2] Nuestro equipo de investigación, INFOSUR, en una trabajo que ha desarrollado entre 2005 y 2008, ha realizado la implantación en máquina mediante el software SMORPH de 4000 verbos del español, 6000 sustantivos, 3000 adjetivos, preposiciones, pronombres, etc.

Hasta el momento ha logrado los siguientes resultados:

a) Delimitación de límites de oraciones (parte de la tesis de Doctorado 2008 de Celina Beltrán bajo la dirección de Gabriel G. Bès)

b) Descripción, formalización y análisis automático del sintagma nominal núcleo (tesis de Maestría 2006, de Andrea Rodrigo, dirigida por Zulema Solana)

c) Descripción, formalización y análisis automático del sintagma verbal núcleo (trabajos de Gabriel Bès y Zulema Solana y tesis doctoral 2010 de Rodolfo Bonino.

d) Descripción, formalización y análisis automático del sintagma adverbial núcleo(tesis doctoral de Andrea Rodrigo)

e) Análisis automático de la puntuación en español (tesis doctoral de Walter Koza)

[3] Aït Mokthar, Salah 1998 L´analyse présintaxique en une seule etape tesis doctoral dirigida por Gabriel G. Bès en el GRIL Université Blaise-Pascal, Francia. Aït-Mokhtar, Salah y Rodrigo Mateos, José Lázaro. 1995 Segmentación y análisis morfológico de textos en español utilizando el sistema SMORPH. SEPLN, 17, 29-41.

[4] Solana, Zulema, Bonino Rodolfo y Valenti, Viviana 2005 Modelización de las fuentes declarativas en una herramienta de análisis y conjugación automáticos de verbos del español en Desarrollo, implementación y uso de modelos para el procesamiento automático de textos (ed. Víctor Castel) Facultad de Filosofía y Letras, UNCUYO

[5] La muestra recoge expresiones de María, Irene y Jakshon (banco de datos CHILDES)

[6] Mariscal, S. 2008 “Early acquisition of gender agreement in the Spanish noun phrase: starting small” en J. Child Lang. 35, 1–29. Cambridge University Press

[7] López Ornat, S.1995 Adquisición de la sintaxis española, Siglo XXI,