Detección de estructuras lingüísticas utilizadas por aprendientes del Español como L2

Claudia Deco, Cristina Bender
Universidad Nacional de Rosario
deco@fceia.unr.edu.ar, bender@fceia.unr.edu.ar

Resumen

En este trabajo se presenta una propuesta para la identificación de estructuras lingüísticas utilizadas por los aprendientes del español como segunda lengua. Para esto se plantea un conjunto de consultas a una base de datos que contiene los textos escritos en español por los aprendientes, previamente etiquetados. El etiquetado se realiza de modo automático con un analizador morfológico (Aït-Mokhtar, 1998) (Aït-Mokhtar, 1995), y la desambiguación se efectúa con información lingüística y se complementa con técnicas estadísticas. Las etiquetas utilizadas corresponden a rasgos y valores definidos en (Solana, 2006). En la teoría de bases de datos existen diversos modelos que pueden utilizarse para representar, almacenar y recuperar la información. La propuesta de este trabajo es utilizar el modelo relacional de bases de datos, aprovechando la versatilidad que brinda este modelo en las implementaciones de las consultas a través del lenguaje de consulta estructurado SQL (Structured Query Language). En este modelo, la información se representa en forma de tablas, donde cada fila corresponde a un elemento dado (en nuestro caso una acepción de una palabra), y cada columna a un atributo descriptivo de la misma (Silberschatz, 2003). Para la implementación se opta por una base de datos relacional, en la que cada palabra del texto se representa en una fila cuyos atributos son la posición de la palabra en el texto y las etiquetas provistas por el analizador (Deco et al., 2008). Este diseño permite extraer listas de secuencias tales como: nombres antecedidos de artículos, nombres antecedidos de artículos con uno o más adjetivos entre ellos, nombres antecedidos de artículos con uno o más adjetivos y adverbios entre ellos, entre otros. Además, es posible contar la cantidad de ocurrencias de este tipo de construcciones a fin de poder obtener estadísticas de su utilización. Con la realización de las consultas, es factible entonces encontrar cuáles son algunas desviaciones comunes realizadas por los aprendientes del español como L2. Además, es posible evaluar si este tipo de desviaciones tienen alguna relación con su lengua materna.

Palabras claves: Base de datos, Detección automática de desviaciones

1. INTRODUCCIÓN

En los aprendientes de español como L2 se presentan diversos tipos de problemas tales como desviaciones verbales, desviaciones de nombres, asignación de número y género diferentes al español estándar, entre otros. Estos problemas han sido tratados en (Solana et al., 2009) (Tramallino, 2009)(Mendez, 2009). Es intención en este trabajo proponer una forma de encontrar estas estructuras en L2 en la formación de los snn, sadvn, etc. utilizando los resultados de los

trabajos mencionados, como ser las salidas de Smorph, MPS, etc. Para esto se propone producir una base de datos que le permita al usuario consultar categorías gramaticales, contextos sintácticos, frecuencia de ocurrencias de construcciones, uso de la puntuación, tiempos y modos verbales, etc. Para la experimentación, se pobló esta base de datos con palabras provenientes de textos producidos por aprendientes de español como L2 que han terminado el nivel inicial y son hablantes de distintas L1 (francés, holandés, alemán e inglés), recopilado por el grupo de investigación INFOSUR de la Universidad Nacional de Rosario.

En (Deco et al., 2008) se propuso el diseño de una base de datos para analizar construcciones en el español estándar. Este diseño es también aplicable para el análisis del español como L2. La carga de la base de datos se realiza a partir de los textos etiquetados. El etiquetado se realiza de modo automático con un analizador morfológico, y la desambiguación se efectúa con información lingüística y se complementa con técnicas estadísticas. Para la implementación se opta por una base de datos relacional, en la que cada palabra del texto se representa en una fila cuyos atributos son la posición de la palabra en el texto y las etiquetas provistas por el analizador. Este diseño permite analizar los problemas mencionados mediante la producción de listados de desviaciones verbales, desviaciones de nombres, asignaciones de género y número diferentes al español estándar, etc. Además, es posible contar la cantidad de ocurrencias de cada tipo de construcciones a fin de poder obtener estadísticas de su utilización.

2. ETIQUETADO MORFOLÓGICO

Para la carga de la base de datos, en primer lugar se efectúa el análisis del corpus de textos, mediante una herramienta que lo segmenta y etiqueta morfológicamente. En este caso se recurre a SMORPH (Aït-Mokhtar, 1998), que tokeniza y efectúa un primer análisis morfológico, sin resolver las ambigüedades. A partir de un texto, como por ejemplo el que se muestra en la Figura 1, este tipo de herramientas genera un archivo etiquetado, tal como el que se muestra en la Figura 2.

También conta este artículo que el asesinator y víctima hacían y la fecha del crimen. Después el crimen el robó su portafolios, rompió una photo y el documento de identitad y escondió debajo un armatorio. La policía localicó a el asesinator después de estudiar muestras de ADN.

Figura 1: Texto de ejemplo para consultas

'También'.
['también', 'EMS','adv'].

'conta'.
['conta', 'EMS','dsverb'].

'este'.
['este', 'EMS','det', 'TDET','dem'].

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

'que'.
['que', 'EMS','rel'].
['que', 'EMS','sub'].

'el'.
['el', 'EMS','det', 'TDET','art'].

'asesinator'.
['asesinator', 'EMS','dsvnom'].

'y'.
['y', 'EMS','cop'].

'víctima'.
['víctima', 'EMS','nom', 'GEN','fem', 'NUM','sg'].

'hacían'.
['hacer', 'EMS','v', 'MODOV','ind', 'PERS','3a', 'NUM','pl', 'TPO','imp].

'y'.
['y', 'EMS','cop'].

'la'.
['el', 'EMS','det', 'TDET','art'].
['lo', 'EMS','cl', 'TPCL','nrfl'].

'fecha'.
['fecha', 'EMS','nom', 'GEN','fem', 'NUM','sg'].

'del'.
['del', 'EMS','contr'].

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

'.'.
['pfp', 'EMS','pun'].

'Después'.
['después', 'EMS','adv'].

'el'.
['el', 'EMS','det', 'TDET','art'].

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

'él'.
['él', 'EMS','pron', 'TPRON','prpers'].

'robó'.
['robar', 'EMS','v', 'MODOV','ind', 'PERS','3a', 'NUM','sg', 'TPO','prets', 'TR','r', 'TC','c1'].

'su'.
['su', 'EMS','det', 'TDET','pos'].

'portafolios'.
['portafolios', 'EMS','nom'].

','.
['cc', 'EMS','coma'].

'rompió'.
['romper', 'EMS','v', 'MODOV','ind', 'PERS','3a', 'NUM','sg', 'TPO','prets', 'TR','r', 'TC','c2'].

'una'.
['una', 'EMS','det', 'TINDF2','indf2a'].

'photo'.
['photo', 'EMS','dsvnom'].

'y'.
['y', 'EMS','cop'].

'el'.
['el', 'EMS','det', 'TDET','art'].

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

'de'.
['de', 'EMS','prep'].

'identitad'.
['identitad', 'EMS','dsvnom'].

'y'.
['y', 'EMS','cop'].

'escondió'.
['esconder', 'EMS','v', 'MODOV','ind', 'PERS','3a', 'NUM','sg', 'TPO','prets', 'TR','r', 'TC','c2'].

'debajo'.
['debajo', 'EMS','adv'].

'un'.
['un', 'EMS','det', 'TINDF1','indf1a'].

'armatorio'.
['armatorio', 'EMS','dsvnom'].

' .'.
['pf', 'EMS','pun'].
['npf', 'EMS','pun'].

'La'.
['el', 'EMS','det', 'TDET','art'].
['lo', 'EMS','cl', 'TPCL','nrfl'].

'policía'.
['policía', 'EMS','nom', 'GEN','_', 'NUM','sg'].

'localicó'.
['localicó', 'EMS','dsverb'].

```
'a'.
[ 'a', 'EMS','prep'].

'el'.
[ 'el', 'EMS','det', 'TDET','art'].

'asesinator'.
[ 'asesinator', 'EMS','dsvnom'].

'después'.
[ 'después', 'EMS','adv'].

'de'.
[ 'de', 'EMS','prep'].

'estudiar'.
[ 'estudiar', 'EMS','v', 'MODOV','infin', 'TR','r', 'TC','c1'].

'muestras'.
[ 'muestra', 'EMS','nom', 'GEN','fem', 'NUM','pl'].

'de'.
[ 'de', 'EMS','prep'].

'ADN'.
[ 'ADN', 'EMS','abr'].

'.'.
[ 'pfp', 'EMS','pun'].
```

Figura 2: Ejemplo de salida del analizador morfológico

Este archivo etiquetado es recorrido por un programa que extrae y vuelca esta información a una base de datos, que luego pueda ser consultada y permita realizar los análisis lingüísticos de interés. A continuación se aclaran las etiquetas utilizadas, que corresponden a rasgos y valores (Solana, 2006).

‘EMS’: etiqueta morfosintáctica.

En el cuadro anterior (Figura 2) aparecen las etiquetas morfosintácticas ‘v’ (verbo), ‘nom’ (nombre), ‘adj’ (adjetivo), ‘adv’ (adverbio), ‘prep’ (preposición), ‘det’ (determinante).

En el *verbo*, los rasgos utilizados se pueden clasificar en dos grupos, por un lado, los relacionados con los valores morfológicos de las terminaciones verbales (modo, tiempo, persona, número), por otro lado, los relacionados con la caracterización del tipo de conjugación y con sus aspectos regulares o irregulares.

Primer grupo:

‘MODOV’: tipo de modo (‘ind’ indicativo, ‘subj’ subjuntivo, ‘infin’ infinitivo, ‘imper’ imperativo), ‘TPO’ tipo de tiempo (‘pres’ presente), ‘PERS’ persona (‘1a’ primera, ‘2a’ segunda, ‘3a’ tercera), ‘NUM’ número (‘sg’ singular, ‘pl’ plural).

Segundo grupo:

‘TC’: tipo de conjugación (‘c1’ primera, ‘c2’ segunda, ‘c3’ tercera).

‘TR’: rasgo que indica si el verbo es o no regular (‘r’ regular, ‘irr’ irregular).

‘TIRR’: tipo de irregularidad (‘hiper’ hiperirregular).

A estas etiquetas se agregan las nuevas etiquetas indicando las desviaciones verbales, de nombre, etc. producidas por las reglas generadas por los investigadores del grupo INFOSUR descriptas en (Solana et al., 2009) (Tramallino, 2009). Estas nuevas etiquetas son la que permiten detectar expresiones en L2.

3. LA BASE DE DATOS *CorpusL2*

En la teoría de bases de datos existen diversos modelos que pueden utilizarse para representar, almacenar y recuperar la información. La propuesta de este trabajo es utilizar el modelo relacional de bases de datos, aprovechando la versatilidad que brinda este modelo en las implementaciones de las consultas a través del lenguaje de consulta estructurado SQL (Structured Query Language). En este modelo, la información se representa en forma de tablas, donde cada fila corresponde a un elemento dado (en nuestro caso una acepción de una palabra), y cada columna a un atributo descriptivo de la misma (Silberschatz, 2003).

La propuesta presentada consiste en una base de datos relacional *CorpusL2*, donde se almacena cada palabra en una o más filas, dependiendo de la cantidad de etiquetas morfosintácticas que posea y se guarda información de la ubicación de cada término dentro del texto a analizar. La información posicional consiste en: una identificación del texto en la que se encuentra la palabra, el número de oración en la que está la palabra y la posición de la palabra dentro de la oración. Esta información permite realizar consultas considerando la adyacencia de los términos, que permita por ejemplo encontrar construcciones del tipo Artículo + Nombre donde no concuerden el género y/o el número. La base de datos *CorpusL2* tiene el siguiente diseño:

CorpusL2(Palabra, NroTexto, NroOración, PosiciónEnOración,
EMS, TC, MODOV, TR, TIRR, TPO, PERS, NUM, GEN, ...)

cuyos atributos son:

Palabra: contiene la palabra como aparece en el texto.

NroTexto: contiene el número o identificación del texto donde se encuentra la palabra.

NroOración: corresponde al número de oración dentro del texto donde se encuentra la palabra.

PosiciónEnOración: es un número que representa la posición de la palabra dentro de la oración.

EMS, TC, MODOV, TR, TIRR, TPO, PERS, NUM, GEN, ...: contienen los valores de las etiquetas correspondientes.

Para el texto de la Figura 1, a partir de la salida del analizador morfológico usado, que se presenta en la Figura 2, se obtiene una instancia en la base de datos CorpusL2. En la Figura 3 se muestra un fragmento de esta instancia.

Palabra	NroTexto	NroOración	PosiciónEn Oración	EMS	TC	MODOV	TR	TIRR	TPO	PERS	GEN	NUM	TDET	Origen
También	1	1	1	adv										
conta	1	1	2	dsverb										
este	1	1	3	det							masc	sg	dem	
artículo	1	1	4	nom							masc	sg		
que	1	1	5	rel										
que	1	1	5	sub										
el	1	1	6	det							masc	sg	art	
asesinator	1	1	7	dsvnom										
y	1	1	8	cop										
víctima	1	1	9	nom							fem	sg		
....														

Figura 3: Fragmento de la instancia de la base de datos CorpusL2

En la Figura 3 se muestra resaltado en verde, una ocurrencia de una desviación verbal, y resaltado en amarillo una ocurrencia de una desviación de nombre.

El algoritmo para la carga de la base de datos a partir de la salida del analizador morfológico es:

Tomar un texto del corpus
Generar el archivo etiquetado mediante una herramienta de análisis morfológico.
Procesar este archivo etiquetado para volcarlo a la base de datos
Leer una línea del archivo de salida etiquetado.
Si esta línea comienza con una palabra
Entonces Insertar una fila en la tabla con la palabra, la información posicional y los valores de las etiquetas correspondientes
Si la línea no comienza con una palabra [1]
Entonces Insertar una fila en la tabla con la palabra de la fila anterior, la nueva información posicional y los nuevos valores de las etiquetas correspondientes
Continuar mientras haya líneas en el archivo
Continuar mientras haya texto en el corpus
Fin.

En el modelo relacional, una consulta se expresa en el lenguaje SQL. Una sentencia de consulta SQL tiene la siguiente sintaxis:

[1] En el archivo etiquetado de la Figura 2, hay líneas que no comienzan con una palabra sino con espacios en blanco, por ejemplo ‘que’. Esto ocurre cuando una palabra tiene más de un etiquetado.

SELECT atributos FROM tabla WHERE condición;

donde

atributos: es la lista de columnas que se desea ver en la respuesta.

tabla: es el nombre de la tabla que contiene los datos, en nuestro caso Corpus.

condición: es un predicado que contiene operadores lógicos Y, O y NO.

A partir del texto de la Figura 1 cargado en la base de datos CorpusL2, a continuación se presentan algunas consultas SQL de ejemplo.

4. EJEMPLOS DE CONSULTAS

Ejemplo 1: Para encontrar las *desviaciones verbales* de los aprendientes de español como L2, se puede realizar la siguiente consulta:

```
SELECT Palabra
FROM CorpusL2
WHERE EMS = ‘dsverb’;
```

Obteniéndose el siguiente resultado para el texto que se está analizando:

conta

localicó

Ejemplo 2: Para encontrar las *desviaciones de nombres* de los aprendientes de español como L2, se puede realizar la siguiente consulta:

```
SELECT DISTINCT Palabra
FROM CorpusL2
WHERE EMS = ‘dsvnom’;
```

Obteniéndose el siguiente resultado para el texto de ejemplo que se está analizando:

asesinator

photo

identitad

armatorio

Notar que el uso de DISTINCT produce que el término *asesinator* que aparece dos veces en el texto resulte en una sola ocurrencia en el listado resultante de la consulta.

Ejemplo 3: Si se desea por ejemplo, encontrar la lista de secuencias de nombres antecedidos de artículos determinantes, que no concuerden en género según el español estándar, la consulta en SQL es la siguiente:

```
SELECT   A.Palabra, N.Palabra
FROM  CorpusL2  A, CorpusL2 N
WHERE A.NroTexto = N.NroTexto
        AND A.NroOración = N.NroOración
        AND N.PosiciónOración - A.PosiciónOración = 1
        AND A.EMS = 'det'
        AND A.TDET = 'art'
        AND N.EMS = 'nom'
        AND A.GEN <> N.GEN;
```

Esta sentencia muestra una lista de artículos determinantes seguidos por nombres. En la condición de búsqueda se pide que las dos palabras estén en el mismo texto, la misma oración, que la resta de sus posiciones dé 1 y que la primera sea un artículo definido y la segunda sea un nombre. La última condición pide que el género del artículo sea distinto del género del nombre. En la Figura 4 se grafica esta resolución.

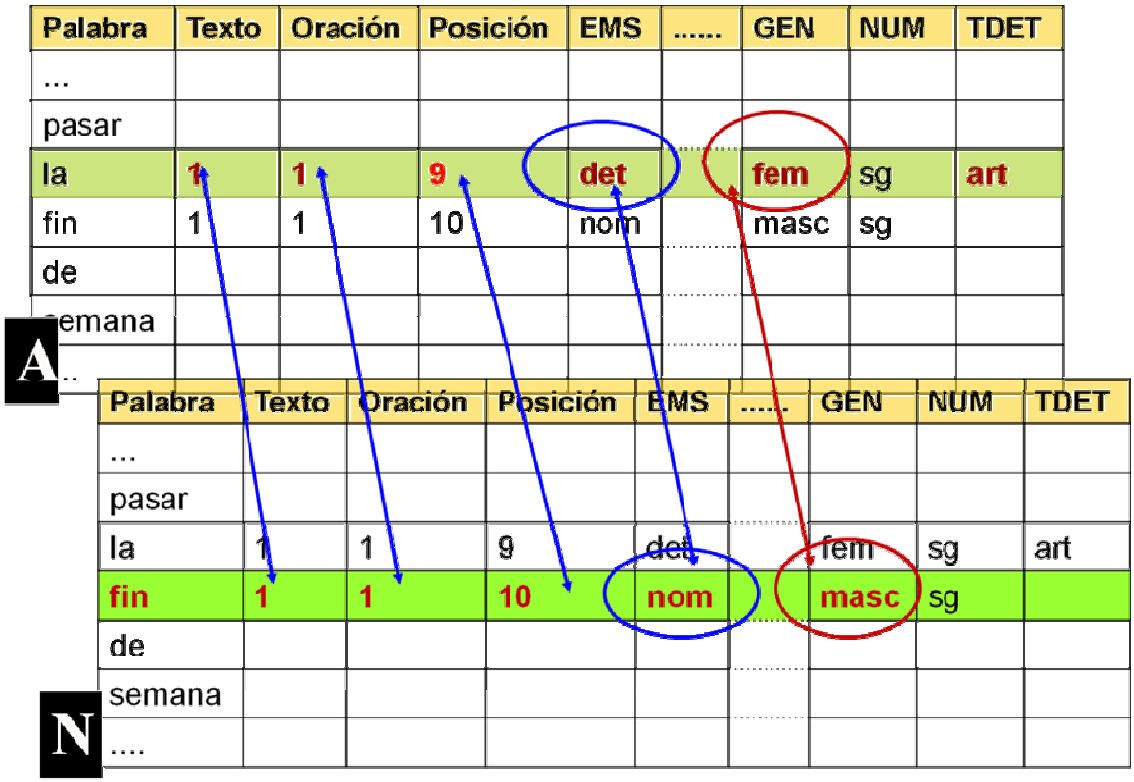

Figura 4: Proceso de la consulta del ejemplo 3

En el texto de ejemplo (Figura 1), no ocurren casos de este tipo. A modo de ejemplo, algunas posibles respuestas a esta consulta serían casos como los siguientes, extraidos de otros textos de aprendientes de español como L2:

la problema

la sistema

los habitaciones

los informaciones

el sangre

la fin

Estos tres ejemplos de consultas presentados, permiten observar la versatilidad del lenguaje de consulta estructurado SQL y del modelo relacional para encontrar solución a los distintos tipos de problemas.

4. CONCLUSIONES

En este trabajo se propuso la utilización de una base de datos relacional que permita efectuar análisis sobre construcciones morfosintácticas distintas al español estándar utilizadas en el idioma español como L2. Para esto, se presentó su diseño, el algoritmo de carga y el uso del lenguaje de consulta SQL para recuperar información. Los casos de uso presentados se ejecutaron sobre una instancia de la base de datos CorpusL2 generada a partir de un fragmento del corpus preparado por el grupo INFOSUR de la Universidad Nacional de Rosario.

Esta propuesta es independiente del analizador morfológico que se utilice. La versatilidad de una base de datos relacional ofrece la ventaja de que con una consulta escrita en SQL es posible recuperar los datos de la forma requerida en cada caso. Es decir, las bases de datos relacionales tienen la capacidad de adaptarse con facilidad y rapidez a diversas funciones. Esto brinda una amplia gama de posibilidades para que los lingüistas puedan analizar diversas construcciones del idioma español mediante la preparación de una consulta adecuada.

Referencias

Aït-Mokhtar, Salah 1998. L'analyse présyntaxique en une seule étape. Tesis doctoral. Universidad Blaise-Pascal/GRIL, Clermont-Ferrand.

Bés, Gabriel; Zulema Solana; Celina Beltrán 2005. “Conocimiento de la lengua y técnicas estadísticas en el análisis lingüístico”. En Víctor Castel (ed.) Desarrollo, implementación uso de modelos para el procesamiento automático de textos. Facultad de Filosofía y Letras, UNCUYO.

Deco, C., Bender, C., Solana, Z. 2008. “Base de datos para el análisis morfosintáctico de un corpus con anotación lingüística”. En Revista INFOSUR. Año 2 Nro 2. Universidad Nacional de Rosario. ISSN 1851 1996. pp 51-60.

Méndez, B. 2009. “Análisis automático de la interlengua: Asignación de género y número diferentes a la lengua estándar en el sintagma nominal núcleo (snn)”. En La interlengua de los aprendientes del español como L2: Aportes de la Lingüística Informática. Rosario: Ediciones Juglaría.

Rodrigo, A. 2009. “El sadvn en la L1 y en la L2”. En La interlengua de los aprendientes del español como L2: Aportes de la Lingüística Informática. Rosario: Ediciones Juglaría.

Silberschatz, A., H. F. Korth 2003. Fundamentos de Bases de Datos, 3 ed., Ed. McGraw-Hill.

Solana, Z. y equipo INFOSUR 2006. Morfología del verbo español, Centro de Estudios de Adquisición del Lenguaje, Facultad de Humanidades y Artes, UNR.

Solana, Z., Beltrán, C., Tramallino, C. 2009. “La implantación en máquina de la interlengua de los aprendientes de español como L2: Los sufijos formadores de nombres”. En La interlengua de los aprendientes del español como L2: Aportes de la Lingüística Informática. Rosario: Ediciones Juglaría.

Tramallino, C. 2009. “Formas verbales irregulares en la interlengua de aprendientes de español como L2”. En La interlengua de los aprendientes del español como L2: Aportes de la Lingüística Informática. Rosario: Ediciones Juglaría.