Análisis automático de ambigüedades en español L2

Automatic analysis of ambiguities in Spanish L2

Prof. Stella Maris Moro
Universidad Nacional de Rosario
Rosario, Argentina
smmoro@yahoo.com.ar

Abstract

Ambiguities constitute a crucial point for the automatic tagging of texts in a natural language. Previously we presented a methodology for the treatment of some of the ambiguities that appear in authentic texts in Spanish and its possible application to the analysis of texts produced by students of Spanish as a second language. Here, we present a proposal for the treatment of these constructions as specific structures of the interlanguage.

Keywords: Ambiguity, Automatic analysis, Disambiguation, Interlanguage, Spanish Second Language.

Resumen

Las ambigüedades constituyen un punto crucial para el etiquetado automático de textos en lenguaje natural. En trabajos anteriores, presentamos una estrategia para el tratamiento de algunas de las ambigüedades que se presentan en textos reales del español y su posible aplicación para el análisis de textos de estudiantes de español como segunda lengua. Aquí presentamos una propuesta para el tratamiento de estas construcciones como estructuras específicas de la interlengua.

Palabras claves: Ambigüedad, Análisis automático, Desambiguación, Interlengua, Español segunda lengua.

1. INTRODUCCION

En trabajos anteriores [1], hemos presentado una estrategia de etiquetamiento de textos en español lengua materna (L1) que permite asignar automáticamente categorías Nombre, Verbo, Adjetivo, etc.

a las cadenas de caracteres. En particular, se abordó la problemática del etiquetado de cadenas de caracteres ambiguas, es decir, cadenas[1] que pueden recibir dos o más etiquetas categoriales[2]:

'joven'.

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

['joven', 'EMS', 'adj', 'GEN', '_', 'NUM', 'sg'].

'trabajo'.

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

['trabajar', 'EMS', 'v', 'MODOV', 'ind', 'PERS', '1a', 'NUM', 'sg', 'TPO', 'pres'].

'sí'.

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

['sí', 'EMS', 'adv'].

Se diseñó un conjunto de rasgos y de reglas de análisis automático que alcanzó una precisión de 99%, con una cobertura del 89,9% en la resolución de este tipo de ambigüedades en función del contexto de aparición de cada ocurrencia.

Posteriormente [2], comenzamos a aplicar esta estrategia en el tratamiento automático de textos en español lengua extranjera (L2). En esos trabajos se registró la existencia de secuencias que resultaban particularmente difíciles de etiquetar, dado que en ellas confluían dos factores: la ambigüedad categorial de las cadenas involucradas y la aparición de ambigüedades particulares en construcciones propias de la interlengua, no posibles en L1. Esta situación daba lugar a una disminución de la efectividad del etiquetado.

[1] Denominamos “cadena” a la sucesión de caracteres que corresponden a lexemas en la lengua natural, y “secuencias” a la sucesión de cadenas que constituyen sintagmas nominales, verbales, adjetivales, adverbiales u oraciones.

[2] Etiquetas:
‘EMS’: Etiqueta Morfo-Sintáctica (a la que se asocian rasgos de Clase de Palabra o de Tipo de Sintagma)
‘GEN’: género
‘NUM’: número
‘MODV’: modo
‘TPO’: tiempo
Rasgos:
‘nom’: nombre
‘v’: verbo
‘ind’: indicativo
‘pres’. presente
‘cl’:clítico
‘encl’: enclítico
‘adj’: adjetivo
‘adv’: adverbio
‘fem’: femenino
‘masc’: masculino
‘_’: rasgo indistinto
‘sg’: singular
‘pl’: plural.
entre otras.

Presentamos aquí una propuesta de análisis automático que permita etiquetar estas secuencias, de tal modo que se logren dos objetivos:

- Evitar el etiquetado erróneo de este tipo de secuencias
- Asignar a estas secuencias una etiqueta que permita reconocer estas secuencias como construcciones propias de la interlengua.

2. METODOLOGÍA DE ETIQUETADO

Utilizamos dos herramientas computacionales:

a.- *SMORPH* (desarrollado por Aït-Mokhtar[3]): autómata de estados finitos que permite lematizar[4] y analizar morfológicamente las cadenas de caracteres, dando como salida la asignación categorial y morfológica correspondiente a cada ocurrencia, de acuerdo con los rasgos que se declaran.

b.- Módulo Post-Smorph (*MPS*), implantado en máquina por Faiza Abbaci[5]: tiene como input la salida de Smorph y, por medio de reglas de recomposición, descomposición y correspondencia que declara el operador, analiza la secuencia de lemas y rasgos gramaticales que se obtiene como output de Smorph.

En el procesamiento de información, ambas herramientas utilizan fuentes declarativas en archivos que son cargados y pueden ser modificados continuamente por el operador.

Para la aplicación de SMORPH utilizamos los archivos de fuentes cargadas por el equipo Infosur, sobre los que se operaron modificaciones en la declaración de los rasgos necesarios para el tratamiento automático de las ambigüedades.

En cuanto a MPS opera con tres tipos de reglas que especifican secuencias posibles de lemas: reglas de **reagrupamiento, descomposición** y **correspondencia** [2].

Tal como presentamos en [2], la metodología de análisis en la que estamos trabajando realiza un análisis por etapas: cada una de ellas articula la información necesaria para las etapas posteriores. Para lograrlo, nuestra propuesta reagrupa los tres tipos de reglas en cuatro conjuntos diferentes:

- **reglas de ejecución preliminar**: incluye reglas de reagrupamiento y de descomposición que facilitan la aplicación de las reglas siguientes.
- **reglas de ejecución primaria**: reglas que analizan las cadenas y secuencias no ambiguas. Incluye reglas de recomposición que operan sobre la formación de sintagmas que incluyen cadenas no ambiguas, y también desambigua cadenas resolubles en primera instancia, tales como ‘el libro’.
- **reglas de postergación**: opera sobre cadenas ambiguas, no resolubles en primera instancia (tales como ‘la amenaza’). Se trata de reglas de correspondencia que asigna nuevas etiquetas que estarán disponibles para las etapas de análisis posteriores.
- **reglas de ejecución secundaria**: operan sobre cadenas y secuencias que han sido postergadas en ejecuciones anteriores, y aprovecha para esto el etiquetamiento de otras secuencias analizadas en etapas anteriores.

Nuestro corpus reúne textos de español L2, producidos por aprendientes de diferentes lenguas maternas, cuyo contacto con la L2 asciende a 2 ó 3 años de estudio, razón por la que identificaremos el estadio de adquisición como intermedio (Ei). Este corpus asciende a un total de 48.000 palabras.

3. ETIQUETADO DEL CORPUS

El primer paso en esta etapa del trabajo fue el etiquetado del corpus de L2 con las reglas ya declaradas para L1.

3.1. Análisis de los resultados del output

La aplicación de Smorph para etiquetar el corpus dio como resultado la detección de **1596 ambigüedades Nombre / Verbo**, del tipo: ‘trabajo’, ‘fuerza’, ‘poder**’, ‘**preguntas’, ‘resumen’, etc. (100%)

Con las reglas declaradas hasta el momento para MPS, las salidas de esta herramienta produjeron la resolución de 1325 de esas ambigüedades. Esto equivale a un **79% de cobertura**. De las cadenas resueltas, a 1257 se asignó la etiqueta correcta, con lo que se obtuvo un **95% de precisión**.

Cadenas ambiguas Nombre – Verbo ‘ambNV’ detectadas: 1596

Cadenas ‘ambNV’ etiquetadas como ‘N’ o ‘V’ a la salida de MPS: 1325

Cadenas ‘ambNV’ etiquetadas adecuadamente como ‘N’ o ‘V’: 1257

Errores: 68

La tabla 1 muestra algunas de las salidas etiquetadas adecuadamente:

Tabla 1: Output de MPS: resolución de ambigüedades

1	'el **estudio**'. ['el estudio', 'EMS', 'SN', 'EMS', 'D+**N**'].
2	'**estudio** el español'. ['estudio el español', 'EMS', 'SV+SN', 'EMS', '**V**+SN'].
3	'**causa** el malentendido'. ['causa el malentender', 'EMS', 'SV+SN', 'EMS', '**V**+SN'].
4	'**Pienso** la **causa**'. ['pienso lo causa', 'EMS', 'SV+SN', 'EMS', '**V**+D+**N**'].

De las 271 ambigüedades N/V (17% del total de ambigüedades detectadas) que no fueron resueltas, la mayoría corresponde a contextos oracionales no contemplados aún en las reglas declaradas en

esta etapa de investigación. Así, por ejemplo, 74 casos (4.8%) son secuencias del tipo ‘pienso que’ y ‘recuerdo que’ o cadenas ambiguas limitadas por signos de puntuación. Este tipo de ocurrencias, que involucran una conjunción subordinante, no han sido incluidas aún en las reglas declaradas en MPS y corresponden a una etapa posterior de nuestra investigación. Sobre el tratamiento de la puntuación, cf. Koza [6].

3.2. Errores en el output

La tabla 2 presenta algunos de los resultados obtenidos como errores en la salida (output) del etiquetado:

Tabla 2: Errores del output de MPS

1	'es'. ['ser', 'EMS', 'SV', 'EMS', 'V']. 'la **mejora** manera'. ['lo mejora manera', 'EMS', 'SV+SN', 'EMS', 'Cl+V+SN'].
2	'tiene'. ['tener', 'EMS', 'SV', 'EMS', 'V']. '**los** cosas'. ['lo cosa', 'EMS', 'SV', 'EMS', 'Cl+V'].
3	'Puede ser uno'. ['poder ser uno', 'EMS', 'SV+SN', 'EMS', 'V+N']. 'de'. ['de', 'EMS', 'prep', 'TPREP', 'prep1']. '**los** razones'. ['lo razones', 'EMS', 'SV', 'EMS', 'Cl+V'].
4	'**la** tema'. ['lo tema', 'EMS', 'SV', 'EMS', 'Cl+V']. 'es'. ['ser', 'EMS', 'ser'].
5	'entre'. ['entre', 'EMS', 'AmbPV']. '**las** programas'. ['lo programa', 'EMS', 'SV', 'EMS', 'Cl+V'].
6	'usando'. ['usar', 'EMS', 'SV', 'EMS', 'Ger']. 'la voz **fuerza**'. ['el voz fuerza', 'EMS', 'SN+SV', 'EMS', 'SN+V'].

Estas estructuras corresponden a diferentes fenómenos de interlengua:

3.2.1. Morfología

El caso (1) presenta la asignación de la desinencia –a a un adjetivo femenino que es invariable. La cadena de caracteres resultante es equivalente a la 3ª pers. sg. del presente de indicativo (y también modo imperativo, 2ª sg.), de modo tal que resulta una "ambigüedad" que es específica de la interlengua.

3.2.2. Asignación de género

Los casos de (2), (3), (4) y (5) se originan por atribuir un género diferente al sustantivo. El sintagma nominal de interlengua resultante se corresponde con sintagmas verbales en L1 y da lugar a una nueva ambigüedad inexistente en L1.

3.2.3. Cambio de categoría léxica

En la ocurrencia de (6), un sustantivo aparece en posición adjetiva. El sujeto utiliza la forma del sustantivo como equivalente al adjetivo. En otras palabras, la secuencia 'fuerza' para este sujeto puede recibir tanto la etiqueta 'N' como 'Adj'.

4. PROPUESTA DE ANÁLISIS AUTOMÁTICO

Dado que lo que hemos detectado son secuencias específicas de un estadio intermedio de adquisición de L2 (Ei), lo que proponemos aquí es declarar una serie de reglas que permitan identificarlas, evitando una asignación errónea de etiquetas.

4.1. Reglas (Ei) declaradas en MPS

La tabla 3 que sigue contiene ejemplos de reglas declaradas en MPS para etiquetar como 'Ei' secuencias como las ejemplificadas en la tabla 2.

Tabla 3: Reglas declaradas en MPS para etiquetamiento de Secuencias Ei

R1: 'EMS', 'ser' + 'EMS', 'SV' → 'EMS', 'Ei'
R2: 'EMS', 'SV' + 'EMS', 'ser' → 'EMS', 'Ei'
R3: 'EMS', 'SV' + 'EMS', 'SV' → 'EMS', 'Ei'
R4: 'EMS', 'prep' + 'EMS', 'SV' → 'EMS', 'Ei'
R5: 'EMS', 'ambPV' + 'EMS', 'SV' → 'EMS', 'Ei'

Obsérvese que

- la regla 1 se aplica a casos como (1), donde el etiquetamiento había p dos SV consecutivos,
- la regla 2 corresponde a casos como (4)
- la regla 3 etiqueta secuencias como las de (2) y (6)[3]
- la 4 es eficiente en casos como (3)
- la número 5 se aplica a (5) donde la secuencia 'entre' es ambigua ('entre' como preposición y 'entre' como forma del verbo 'entrar').

5. RESULTADOS DE LA APLICACIÓN DE REGLAS (Ei)

La aplicación de las reglas precedentes devuelve el siguiente etiquetado como output:

Tabla 4: Reconocimiento de Secuencias 'Ei'

1	'es la mejora manera'. ['ser lo mejora manera', 'EMS', 'Ei'].
2	'tiene los cosas'. ['tener lo cosa', 'EMS', 'Ei'].
3	'Puede ser uno'. ['poder ser uno', 'EMS', 'SV+SN', 'EMS', 'V+N']. 'de los razones'. ['de lo razones', 'EMS', 'Ei'].
4	'la tema es'. ['lo tema ser', 'EMS', 'Ei'].
5	'entre las programas'. ['en lo programa', 'EMS', 'Ei'].
6	'usando la voz **fuerza**'. ['usar el voz fuerza', 'EMS', 'Ei'].

6. ANÁLISIS DE RESULTADOS

Como puede observarse en la tabla 4, el output obtenido presenta la etiqueta 'Ei' para aquellas secuencias que habíamos identificado como específicas de la interlengua. Ahora bien, al efectuar la

[3] Quedan excluidas secuencias SV + SV como la de "la casa que **construí se derrumbó**" gracias a la declaración de reglas de reconocimiento de subordinadas (tales como: 'EMS' 'sub' + 'EMS' 'SV' → 'OSub', donde 'sub' equivale a 'subordinante' y 'OSub' a la etiqueta de oración subordinada).

búsqueda de todas las secuencias que el autómata identifica ahora como ‘Ei’ obtenemos salidas como las presentadas en la tabla 5:

Tabla 5: Secuencias ‘Ei’ detectadas

7	'sin **basa**'. ['sin basar', 'EMS', 'Ei'].
8	'de juego'. ['de juego', 'EMS', 'SP', 'EMS', 'P+N']. 'de **role**'. ['de rolar', 'EMS', 'Ei'].
9	’en la **secundaría'**. ['en lo secundar', 'EMS', 'Ei'].
10	'de **Medico'**. ['de medicar', 'EMS', 'Ei'].
11	'de **practica'**. ['de practicar', 'EMS', 'Ei'].
12	'en **la trabaja'**. ['en lo trabajar', 'EMS', 'Ei'].
13	'tienen que seguir **la** modelo'. ['tener que seguir lo modelo', 'EMS', 'Ei'].
14	'es **podemos'**. ['ser poder', 'EMS', 'Ei'].
15	'podemos **aprendemos'**. ['poder aprender', 'EMS', 'Ei']. 'español'. ['español', 'EMS', 'adj', 'GEN', 'masc', 'NUM', 'sg'].
16	'hablar'. ['hablar', 'EMS', 'Inf']. '**con** pensando'. ['con pensar', 'EMS', 'Ei'].
17	'en **nos** colegio'. ['en lo colegiar', 'EMS', 'Ei'].

Los ejemplos (7) y (8) presentan una morfología específica de interlengua, similar a lo apuntado en §3.2.1., y las terminaciones –a y –e asignadas respectivamente provocan el surgimiento de una ambigüedad que no está presente en L1.

En cuanto a ejemplos como (9), (10) y (11), aparece un nuevo fenómeno, relacionado con la grafía (presencia / ausencia de tilde). La falta de tilde en 'practica' y 'Medico', y su presencia en 'secundaría' hacen que, en la escritura de este sujeto, tales cadenas de caracteres sean idénticas ya sea que las utilice como sustantivo o verbo[4]. Estas cadenas, que el autómata habría etiquetado como preposición + verbo, a partir de la implantación de reglas específicas de la interlengua son reconocidas como 'Ei'.[5]

El caso (12) combina morfología (§3.2.1.) y asignación de género (§3.2.3.) específicas de la interlengua, mientras que en (13) también se asigna género femenino a un sustantivo masculino.

(14) y (15) ponen en evidencia procesos de morfología verbal particulares de la adquisición de L2, al asignar morfología personal a verbos en contextos de aparición de infinitivos.

El ejemplo de (16) tiene una naturaleza lingüística totalmente diferente, pues se produce a nivel de la configuración sintáctica al anteponer una preposición al gerundio 'pensando'.

Por último, (17) presenta un pronombre personal en posición de adjetivo ('nos' por 'nuestros'). Se trata de un caso particular de ambigüedad específica de la interlengua, semejante a las presentadas en 3.2.1., pero involucrando a la categoría Pronombre.

7. CONCLUSIONES Y PERSPECTIVAS DE DESARROLLO

Las reglas presentadas constituyen una aproximación primaria a la detección de estas secuencias en L2 a través de las herramientas utilizadas, y se encuentra aún en fase experimental. Si bien se ha constatado la aplicación de las reglas declaradas para la resolución de las secuencias presentadas, se trata de un abordaje parcial, y nuevas reglas deberán declararse para otras secuencias de mayor complejidad.

Sin embargo, las 5 reglas de 'Ei' declaradas hasta aquí han resultado efectivas, pues posibilitaron la identificación automática de secuencias de interlengua de forma de limitar el tiempo de reconocimiento que demanda esta tarea en lápiz y papel.

Por otra parte, el análisis de estos ejemplos, detectados por el autómata sin necesidad de leer 'manualmente' todo el corpus, permitió ampliar el espectro de fenómenos de interlengua a casos no previstos inicialmente.

Muchos aspectos relevantes para el análisis de la interlengua quedan esbozados aquí. Entre ellos:

- la posibilidad de efectuar un relevamiento estadístico de la aparición de cada uno de los fenómenos detectados;
- la frecuencia de aparición de cada uno de los fenómenos podría permitir determinar diferentes estadios de adquisición (Ei1, Ei2, etc.)

[4] Algo similar ocurría en §3.2.3., aunque aquí el fenómeno se debe específicamente a la presencia o ausencia de tilde.
[5] En lo referido a la tildación, podemos prever un fenómeno similar en sujetos que están adquiriendo el español como L1.

- la aparición de cadenas correspondientes a una categoría (por ej., pronombre o verbo) en posiciones propias de otra (adjetivo o nombre respectivamente) puede constituir un punto de abordaje para el análisis lingüístico de cada estadio de adquisición.

A partir de aquí, las implicancias para el diseño de estrategias de profesores de español L2 parecen abrir diferentes vías de desarrollo, no sólo como medio de detección de secuencias de interlengua, sino también como punto de partida para la intervención didáctica posterior, anclada en el análisis lingüístico de estas ocurrencias.

Referencias

[1]Moro, S.M. (2008) "Análisis automático de ambigüedades en español: las categorías 'nombre' y 'verbo'", en Infosur 2:15-26 http://www.infosurrevista.com.ar.

[2] Moro, S.M. (2009) "Análisis automático de producciones de estudiantes japoneses de español L2: la resolución de ambigüedades", en Infosur 3: 73-82. http://www.infosurrevista.com.ar.

[3] Aït-Mokthar S. (1998) L´analyse présintaxique en une seule etape. Tesis doctoral dirigida por Gabriel G. Bès en el GRIL, Université Blaise-Pascal, Francia, 1998.

[4] Lema: cadena que designa por convención toda la serie de variantes morfológicas de un lexema. Por ej.: 'libro' es el lema que designa el paradigma nominal: 'libro / libros', mientras que 'librar' es la denominación del paradigma verbal 'libro / libras / librás / libra /…'

[5] Abbaci F. Développement du Module Post-Smorph. Memoria del DEA de Linguistique et Informatique. Universidad Blaise-Pascal/GRIL, Clermont-Fd, 1999.

[6] Koza (2008) "Análisis automático de textos: Reconocimiento de incisos", en Infosur 2:73-84; (2009) "Análisis automático de textos: Reconocimiento de construcciones dicendi", 3: 95-104 http://www.infosurrevista.com.ar.