Análisis Automático de Producciones de Estudiantes Japoneses de Español L2: la Resolución de Ambigüedades

AUTOMATIC ANALYSIS OF TEXTS PRODUCED BY JAPANESE STUDENTS OF SPANISH L2: THE RESOLUTION OF AMBIGUITIES

Prof. Stella Maris Moro
GRUPO INFOSUR UNR
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 model for the treatment of some of the ambiguities that appear in authentic texts in Spanish, particularly those referred to the 'noun' and 'verb' categories. Here, we propose its possible application to the analysis of texts produced by students of Spanish as a second language, evaluate the results obtained and anticipate some possible projections of this work.

Keywords: Ambiguity, Automatic analysis, Disambiguation, POS Tagging, Spanish Second Language.

Resumen

Las ambigüedades constituyen un punto crucial para el etiquetado automático de textos en lenguaje natural. En trabajos anteriores, presentamos un modelo para el tratamiento de algunas de las ambigüedades que se presentan en textos reales del español, en particular, las referidas a las categorías 'nombre' y 'verbo'. Aquí proponemos su posible aplicación para el análisis de textos de estudiantes de español como segunda lengua, evaluamos los resultados alcanzados y prevemos algunas proyecciones posibles de este trabajo.

Palabras claves: Ambigüedad, Análisis automático, Desambiguación, Etiquetado gramatical, español segunda lengua.

1. INTRODUCCION

El análisis de textos a través de la utilización de etiquetadores automáticos constituye en la actualidad una herramienta crucial tanto en la implementación de traductores, correctores ortográficos y sintácticos, y buscadores, como en el desarrollo de la investigación lingüística basada en el tratamiento estadístico de grandes corpora.

Sin embargo, las ambigüedades continúan siendo un problema crucial para los etiquetadores. Por un lado, los programas que operan en forma estadística presentan un margen de error importante en estructuras bastante simples. Por otro, aquellos que requieren de entrenamiento con textos etiquetados implican un esfuerzo manual que no se traduce en una minimización efectiva del margen de error.[1]

En trabajos anteriores [2], hemos presentado una modelización que alcanzó un alto grado de precisión y cobertura en la resolución de algunos tipos de ambigüedades: adjetivo – sustantivo (ej.: joven, médico), sustantivo – verbo (ej.: deber, trabajo), sustantivo – adverbio (más, sí).

La utilización de esta herramienta provee un nuevo recurso en el campo del análisis de producciones de textos de interlengua.

En este trabajo, presentamos la modelización propuesta inicialmente para el etiquetamiento de textos en español redactados por hablantes nativos, revisamos luego los alcances que, tendría en el análisis de producciones de estudiantes japoneses que se encuentran en un estadio intermedio de adquisición del español como segunda lengua y por último consideramos las proyecciones que este tipo de perspectiva puede implicar para la enseñanza del español como segunda lengua.

2. MODELIZACIÓN

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.[2]

MPS opera con tres tipos de reglas que especifican secuencias posibles de lemas:

- reglas de **reagrupamiento**: una secuencia dos o más cadenas se reescriben como una secuencia de mayor nivel. Por ej.,

  Artículo + Nombre → SN.

- reglas de **descomposición**: una cadena se reescribe como secuencia de dos o más cadenas :

  Contracción → Preposición + Determinante

- reglas de **correspondencia**: una cadena o secuencia con un rasgo determinado se reescribe con otro rasgo:

  Artículo → Determinante

  SN → sn

El modelo en el que estamos trabajando realiza un análisis por etapas: cada una de ellas articula la información necesaria para las etapas posteriores. Para lograrlo, nuestra modelización reagrupa los tres tipos de reglas en cuatro conjuntos diferentes:

- **reglas de ejecución preliminar**: incluye reglas de reagrupamiento y de descomposición que son imprescindibles para todo análisis ulterior. Incluye:

✓ Reglas de descomposición de **contracciones**, que permite disponer de la etiqueta 'Det' para el reconocimiento de SN en la etapa siguiente.

Contracc → Prep + Det

✓ Reglas de recomposición de **SV**, que reagrupa las frases verbales, desambiguando cadenas ambiguas como "deber" (N/V) "regular" (V/Adv) o "pasado" (N/A/Ppio) que aparecen en secuencias verbales.

VerboAux + 'que' + 'Infinitivo' → **SV**

- **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.

ambNV → AmbNV

Así, por ej., la cadena 'amenaza' es etiquetada como 'ambNV' (ambiguo Nombre/Verbo) por Smorph. Este rasgo es considerado por MPS y en esta etapa reescrito como 'AmbNV', rasgo que será el input de las reglas de ejecución secundaria.

- **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.

Hasta ahora, el modelo se aplicó al tratamiento de cadenas ambiguas Nombre / Verbo, Nombre / Adjetivo, y Nombre / Adverbio. Se evaluó la efectividad de esta modelización en un corpus de 10190 palabras con el siguiente resultado:

Precisión: 96%

Cobertura: 89.8 %.[2]

3. APLICABILIDAD DEL MODELO EN EL ANÁLISIS DE L2

Consideramos que la modelización que estamos construyendo tiene aplicaciones que no se limitan al análisis y etiquetamiento de textos del español estándar. Nos interesa aquí testear su aplicabilidad a textos de interlengua.

Para ello trabajamos con un corpus de **203 textos** en español producidos por estudiantes japoneses de español como lengua extranjera. Todos los sujetos tienen entre dos y tres años de estudio de la L2. El corpus asciende a un total de **46.000 palabras**.

La salida de smorph, primer paso del etiquetamiento, registra un total de **1542 ambigüedades Nombre / Verbo**, del tipo: 'trabajo', 'fuerza', 'poder**', '**preguntas', resumen, etc.

En la siguiente tabla se ofrecen algunos ejemplos (se han simplificado en parte los rasgos para permitir una visualización más simple de las salidas). La negrita resalta las cadenas ambiguas

Nombre / Verbo (etiqueta ‘TAMB’, rasgo ‘ambNV’)[1]:

Tabla 1: Output *Smorph*: detección de ambigüedades[2]

1	'no'. ['no', 'EMS','adv',]. 'puede'. ['poder', 'EMS','v']. 'comprender'. ['comprender', 'EMS','v']. 'unas'. ['unas', 'EMS','det', 'TAMB','ambDV']. ['unir', 'EMS','v']. **'preguntas'.** ['pregunta', 'EMS','nom', 'GEN','fem', 'NUM','pl', 'TAMB',**'ambNV'**]. ['preguntar', 'EMS','v'].
2	'me'. ['lo', 'EMS','cl']. 'da'. ['dar', 'EMS','v']. **'ganas'.** ['gana', 'EMS','nom', 'GEN','fem', 'NUM','pl', 'TAMB',**'ambNV'**]. ['ganar', 'EMS','v']. 'de'. ['de', 'EMS','prep']. 'estudiar'. ['estudiar', 'EMS','v']. 'lo'. ['lo', 'EMS','encl'].

[1] Para la etiqueta ‘TAMB’ (Tipo de Ambigüedad) se declararon rasgos ‘ambNV’ para las cadenas ambiguas Nombre/Verbo (del tipo ‘camino’), ‘ambPV’ para ambigüedades Preposición/Verbo (‘sobre’), etc.

[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
Rasgos:
‘nom’: nombre
‘v’: verbo
‘cl’:clítico
‘encl’: enclítico
‘adj’: adjetivo
‘adv’: adverbio
‘fem’: femenino
‘masc’: masculino
‘sg’: singular
‘pl’: plural.

3	'cuando'. ['cuando', 'EMS','rel']. 'tengamos'. ['tener', 'EMS','v']. **'dudas'.** ['duda', 'EMS','nom', 'GEN','fem', 'NUM','pl', **'TAMB','ambNV'**]. ['dudar', 'EMS','v'].
4	'es'. ['ser', 'EMS','ser']. **'doble'.** ['doble', 'EMS','adj', 'GEN','_', 'NUM','sg']. ['doble', 'EMS','nom', 'GEN','masc', 'NUM','sg', **'TAMB','ambNV'**]. ['doblar', 'EMS','v']. **'trabajo'.** ['trabajo', 'EMS','nom', 'GEN','masc', 'NUM','sg', **'TAMB','ambNV'**]. ['trabajar', 'EMS','v'].
5	'en'. ['en', 'EMS','prep']. **'corro'.** ['corro', 'EMS','nom', 'GEN','masc', 'NUM','sg', **'TAMB','ambNV'**]. ['correr', 'EMS','v'].
6	'la'. ['el', 'EMS','det', 'GEN','fem', 'NUM','sg']. ['lo', 'EMS','cl', 'GEN','fem', 'NUM','sg']. 'gente'. ['gente', 'EMS','nom', 'GEN','fem', 'NUM','sg']. **'cerca'.** ['cerca', 'EMS','adv', 'TAMB','ambVAv']. ['cerca', 'EMS','nom', 'GEN','fem', 'NUM','sg', **'TAMB','ambNV'**]. ['cercar', 'EMS','v']. 'la'. ['el', 'EMS','det', 'GEN','fem', 'NUM','sg']. ['lo', 'EMS','cl', 'GEN','fem', 'NUM','sg']. 'pareja'. ['pareja', 'EMS','nom', 'GEN','fem', 'NUM','sg'].
7	'el'. ['el', 'EMS','det', 'GEN','masc', 'NUM','sg', 'TDET','art']. **'estudio'.** ['estudio', 'EMS','nom', 'GEN','masc', 'NUM','sg', **'TAMB','ambNV'**]. ['estudiar', 'EMS','v'].
8	**'estudio'.** ['estudio', 'EMS','nom', 'GEN','masc', 'NUM','sg', **'TAMB','ambNV'**]. ['estudiar', 'EMS','v'].

	'el'. ['el', 'EMS','det', 'GEN','masc', 'NUM','sg', 'TDET','art']. 'español'. ['español', 'EMS','adj', 'GEN','masc', 'NUM','sg'].
9	**'causa'.** ['causa', 'EMS','nom', 'GEN','fem', 'NUM','sg', 'TAMB','ambNV']. ['causar', 'EMS','v']. 'el'. ['el', 'EMS','det', 'GEN','masc', 'NUM','sg', 'TDET','art']. 'malentendido'. ['malentender', 'EMS','v', 'MODOV','part', 'GEN','masc', 'NUM','sg'].
10	**'Pienso'**. ['pienso', 'EMS','nom', 'GEN','masc', 'NUM','sg', **'TAMB','ambNV'**]. ['pensar', 'EMS','v']. 'la'. ['el', 'EMS','det', 'GEN','fem', 'NUM','sg', 'TDET','art']. ['lo', 'EMS','cl', 'GEN','fem', 'NUM','sg']. **'causa'.** ['causa', 'EMS','nom', 'GEN','fem', 'NUM','sg', **'TAMB','ambNV'**]. ['causar', 'EMS','v'].

3.1. Análisis de los resultados del modelo.

De las 1542 cadenas 'ambNV' presentes en el corpus detectadas por Smorph, la modelización propuesta **resuelve 1282**, lo que da un **83% de cobertura**.

De esas 1282, el número de ambigüedades bien resueltas asciende a **1215** cadenas, con lo cual la **precisión** alcanzada por el modelo en el análisis de textos de español L2 es de **78.8%**.

A continuación se insertan las secuencias del archivo de salida de MPS, correspondientes a los ejemplos presentados en la tabla 1. La negrita resalta las cadenas etiquetadas como ambiguas Nombre / Verbo ('ambNV') por Smorph, y su etiquetado por MPS como N o V, según corresponde en cada secuencia.

Tabla 2: Output *MPS*: resolución de ambigüedades

1	'no'. ['no', 'EMS', 'Adv']. 'puede comprender'. ['poder comprender', 'EMS', 'SV', 'EMS', 'V+Vinfin']. 'unas **preguntas**'. ['unas pregunta', 'EMS', 'SN', 'EMS', 'D+**N**'].
2	'me da **ganas**'. ['lo dar gana', 'EMS', 'SV+SN', 'EMS', 'V+**N**']. 'de estudiar'. ['de estudiar', 'EMS', 'SP', 'EMS', 'P+Vinf'].

	'lo'. ['lo', 'EMS', 'encl'].
3	'cuando'. ['cuando', 'EMS', 'rel']. 'tengamos **dudas**'. ['tener duda', 'EMS', 'SV+SN', 'EMS', 'V+**N**'].
4	'es'. ['ser', 'EMS', 'ser']. '**doble trabajo**'. ['doble trabajo', 'EMS', 'SN', 'EMS', '**A+N**'].
5	'en **corro'**. ['en corro', 'EMS', 'SP', 'EMS', 'P+**N**'].
6	'la gente **cerca**'. ['el gente cerca', 'EMS', 'SV+SN', 'EMS', 'SN+**V**']. 'la pareja'. ['el pareja', 'EMS', 'SN', 'EMS', 'D+N'].
7	'el **estudio**'. ['el estudio', 'EMS', 'SN', 'EMS', 'D+**N**'].
8	'**estudio** el español'. ['estudio el español', 'EMS', 'SV+SN', 'EMS', '**V**+SN'].
9	**'causa** el malentendido'. ['causa el malentender', 'EMS', 'SV+SN', 'EMS', '**V**+SN'].
10	'**Pienso** la **causa'**. ['pienso lo causa', 'EMS', 'SV+SN', 'EMS', '**V**+D+**N**'].

Queda un remanente de 264 ambigüedades N/V (17% del total de ambigüedades detectadas) que no fueron resueltas y quedan señaladas en el output como cadenas ambiguas Nombre / Verbo 'AmbNV' (ej: 'pienso') o como sintagmas ambiguos SN / SV 'AmbSNSV' (del tipo 'las causas'). En general, corresponden 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'. 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.

3.2. Errores en el output

Durante el control del archivo de output de MPS, se detectaron **67 errores** de etiquetamiento (4.5%). de los cuales al menos 10 se deben a estructuras propias de la interlengua, y no a la operatividad de las reglas declaradas. Puede observarse básicamente recurrencia en el tipo de

errores que aparece en la interlengua, y que permite clasificar el tipo de mecanismos gramaticales que motivaron el error de etiquetamiento.

Las tablas 3 y 4 muestran algunas de las secuencias gramaticalmente mal formadas, y el etiquetamiento que se obtuvo a partir de ellas con las reglas declaradas para L1.

3.2.1. Errores de concordancia

Tabla 3: Output *MPS*: etiquetamiento de secuencias mal formadas por concordancia

1	'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	'son'. ['son', 'EMS', 'SV', 'EMS', 'V']. **'claras'**. ['claro', 'EMS', 'adj', 'GEN', 'fem', 'NUM', 'pl']. **'la** objeto'. ['lo objeto', 'EMS', 'SV', 'EMS', 'Cl+V'].
4	'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'].
5	**'la** tema'. ['lo tema', 'EMS', 'SV', 'EMS', 'Cl+V']. 'es'. ['ser', 'EMS', 'ser'].
6	'entre'. ['entre', 'EMS', 'AmbPV']. **'las** programas'. ['lo programa', 'EMS', 'SV', 'EMS', 'Cl+V'].
7	'de'. ['de', 'EMS', 'prep', 'TPREP', 'prep1']. **'los** razones'. ['lo razones', 'EMS', 'SV', 'EMS', 'Cl+V'].

Evidentemente, la frecuencia de este tipo de estructuras parece señalar una tendencia en los textos de ELE2 de hablantes cuya lengua materna es el japonés, y constituir un punto a analizar en la interlengua.

3.2.2. Errores de morfología

Tabla 4: Output *MPS*: etiquetamiento de cadenas con errores de morfología

1	'usando'. ['usar', 'EMS', 'SV', 'EMS', 'Ger']. 'la voz **fuerza**'. ['el voz fuerza', 'EMS', 'SV+SN', 'EMS', 'SN+V']. 'y'. ['y', 'EMS', 'cop']. 'el gesto grande'. ['el gesto grande', 'EMS', 'SN', 'EMS', 'D+N+A'].
2	'para comprender'. ['para comprender', 'EMS', 'SP', 'EMS', 'P+Vinf']. 'la estructura **gramática**'. ['lo estructura gramática', 'EMS', 'SV+SN', 'EMS', 'Cl+V+SN'].

Estos dos ejemplos presentan un caso particular de **utilización ambigua de secuencias no ambiguas** en español: el productor del texto utiliza ‘fuerza’ y ‘gramática’ como ‘adj’. Otro elemento que podrá resultar productivo en el trabajo con este tipo de corpus.

3.2.3. Errores de grafía

En los ejemplos transcriptos en la tabla 5, pueden analizarse los errores de etiquetamiento ocasionados por secuencias que corresponden a palabras inexistentes (‘mi’) o a bien escritas con una grafía que corresponde a otra cadena, inadecuada para la secuencia.

Tabla 5: Output *MPS*: etiquetamiento de cadenas con errores de grafía

1	'la basa'. ['lo basar', 'EMS', 'SV', 'EMS', 'Cl+V'].
2	'las'. ['el', 'EMS', 'det', 'GEN', 'fem', 'NUM', 'pl']. ['lo', 'EMS', 'cl', 'GEN', 'fem', 'NUM', 'pl']. 'parablas'. ['parablas' , mi]. 'largas'. ['largar', 'EMS', 'SV', 'EMS', 'V'].

4. PERSPECTIVAS DE DESARROLLO

Las proyecciones de este tipo de modelización en el análisis de la interlengua son amplias, y aún estamos en etapa de reconocer hasta dónde podemos llegar por este camino.

Sin embargo, podemos vislumbrar que estamos en el principio de lo que puede llegar a ser una herramienta de gran utilidad para la enseñanza del español como L2.

Una de las proyecciones que aparecen claramente viables es el abordaje de la concordancia, que aparece en los textos analizados como una problemática propia de la interlengua. La declaración de reglas de formación de SN que contengan restricciones de género y número deben dar resultados directos en este aspecto.

La utilización de cadenas correspondientes a una categoría (o rasgo categorial: 'N', 'Adj', etc) en posiciones que corresponden a otra, tales como las presentadas en la tabla 4, pueden dar lugar a un interesante análisis acerca de la distribución categorial en la interlengua.

Finalmente, archivos de salida de esta naturaleza permitirán establecer con rapidez y en forma estadística índices relevantes en la determinación del tipo de construcciones gramaticales predominantes en cada estadio de adquisición de la L2

Proyecciones posteriores del análisis podrán volcarse en el diseño de estrategias útiles para el profesor de español como L2.

Referencias

[1] Para un análisis del rendimiento de algunas herramientas on line en el tratamiento de textos de aprendices de español como L2, con alusiones al problema de las ambigüedades, ver Moro, S.M. "Evaluación de herramientas informáticas para el tratamiento de textos de aprendices de español como L2", en AA.VV, *Recursos informáticos para el tratamiento lingüístico de textos*, Grupo Infosur, UNR, Ed. Juglaría, Rosario, 2008.

[2] Cf. Moro, S.M. "Análisis automático de ambigüedades en español: las categorías 'nombre' y 'verbo'", en *Infosur* 2:15-26, Agosto 2008, 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.