# Análisis automático de ambigüedades en español: las categorías 'nombre' y 'verbo'

**Automatic analysis of ambiguities in Spanish: the 'noun' and 'verb' categories**

**Stella Maris Moro**
Facultad de Humanidades y Artes, 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. Here we propose a model for the treatment of some of the ambiguities that appear in authentic texts in Spanish, particularly those referred to the categories 'noun' and 'verb'. The model is based on architecture of rules declared in relation to the sentence context that avoids both the manual labeling of training texts and their statistical treatment, in order to minimize the error margin that these strategies produce. We evaluate the results obtained and anticipate some possible projections of this work.

**Keywords**: Ambiguity, Automatic analysis, Disambiguation, POS Tagging, Spanish.

**Resumen**

Las ambigüedades constituyen un punto crucial para el etiquetado automático de textos en lenguaje natural. Proponemos aquí 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'. El modelo se basa en una arquitectura de reglas declaradas en relación con el contexto oracional, que evita tanto el etiquetado manual de textos de entrenamiento como la operatoria estadística, con el fin de minimizar el margen de error que producen estas estrategias. Evaluamos los resultados alcanzados y prevemos algunas proyecciones posibles de este trabajo.

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

## 1. INTRODUCCION

El tratamiento de las ambigüedades constituye un punto crucial en el análisis automático de textos en lenguaje natural. Herramientas que operan tareas tales como traducción, corrección ortográfica y gramatical, búsqueda de información, tratamiento estadístico, etc., involucran el etiquetado automático y requieren de una asignación precisa de etiquetas gramaticales.

Sin embargo, las ambigüedades continúan siendo uno de los problemas más difíciles de resolver. 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 un entrenamiento previo con textos etiquetados implican un esfuerzo manual que no se traduce en una minimización efectiva del margen de error. Esto se hace muy notorio en español, lengua para la cual actualmente se adaptan los modelos aplicados al inglés, adecuaciones que presentan aún un grado de desarrollo insuficiente.

Dada la complejidad de la problemática, nos proponemos presentar parte de una modelización posible para el tratamiento de algunas ambigüedades que involucran las categorías ‘nombre’ y verbo’ en español.

En primer término circunscribiremos la problemática a algunos tipos de ocurrencias. Presentaremos luego la herramienta informática utilizada y el autómata propuesto. Analizaremos éste en dos etapas: una referida a los nombres, y la otra, complementaria, en relación con los verbos. Por último, evaluaremos los resultados obtenidos y presentaremos algunas posibles proyecciones de este trabajo.

## 2. RECORTE DE CASOS

El tratamiento automático de textos opera a partir de expresiones (E) de la lengua natural que se traducen a lenguaje de máquina como cadenas de caracteres en código ASCII. A partir de un conjunto de reglas declarativas, el programa reconoce y segmenta estas cadenas y luego les asigna una interpretación (I) o etiqueta de rasgos (morfológicos o sintácticos).

En el caso de las ambigüedades, el programa asigna dos o más interpretaciones, de las cuales sólo una resulta adecuada. Por ej.:

E: ‘canto’

se interpretará automáticamente como:

$I_a$: [‘canto’, sustantivo masculino singular]
$I_b$: [‘canto’, verbo presente indicativo 1ª persona singular]

Sin embargo, sólo una de estas interpretaciones es válida en secuencias de cadenas [1]; así en

‘el canto de los pájaros’ ó ‘canto muy mal’

‘canto’ es, respectivamente, ‘sustantivo’ y ‘verbo’.

En español, existe un rango muy amplio de ambigüedades en la asignación de las categorías correspondientes a las clases de palabras [2].

Tabla 1: Tipos de Ambigüedades

| Tipo de Ambigüedad | Cadenas |
|---|---|
| **N / A** | hueco |
| **N / V** | amenaza |
| **Pr / D** | la |
| **N / V inf** | poder |
| **N / V pp** | bebida |
| **A / V** | completa |
| **N / A / V** | presente |
| **A / Adv** | mucho |
| **A / Adv/ V** | regular |
| **N / Adv / V** | cerca |
| **N / Intj / V** | vale |
| **Pr / V** | consigo |

| Tipo de Ambigüedad | Cadenas |
|---|---|
| **Cj / V** | como / sino |
| **D / V** | una |
| **P / V** | entre / para |
| **P / N / V** | sobre |
| **P / N/ A / Adv / V** | bajo |
| **Adv / A / N** | mal |

La mayor parte de estas ambigüedades involucran a la categoría N y/o V [3], de modo que, como veremos, la declaración de reglas que permitan analizar en forma no ambigua estas categorías traerán como consecuencia efectos sobre otras.

En este trabajo nos circunscribiremos a los dos primeros tipos de ambigüedades: N / A y N / V, y analizaremos las implicancias con relación a otras.

## 3. HERRAMIENTAS INFORMÁTICAS

Utilizamos dos herramientas computacionales:

a.- *SMORPH* (desarrollado por Aït-Mokhtar [4]) permite 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 por Faiza Abbaci [5], tiene como input la salida de Smorph. A partir de reglas de recomposición, descomposición y correspondencia que se declaran, analiza la cadena de lemas que se obtiene como salida de Smorph.

Ambas herramientas son de código abierto, ya que en el procesamiento de información utilizan como fuente archivos que son cargados y pueden ser modificados continuamente por el operador.

Las fuentes declarativas de Smorph están constituidas por 5 archivos:

- **ascii.txt**: se declaran códigos ascii específicos tales como los separadores de oración y de párrafo.
- **rasgos.txt**: incluye etiquetas de rasgos morfológicos a aplicar en el análisis de las cadenas de caracteres, con sus posibles valores, por ej:
  EMS {nombre, verbo,…}
  GÉNERO {femenino, masculino, neutro}
- **term.txt**: carga de las diferentes terminaciones que cada lema puede presentar en su derivación morfológica.
  -o, -as, ás, -a, -amos, -áis, -an
- **entradas.txt**: listado de lemas y los modelos correspondientes de derivación.
  casar v1
- **modelos.txt:** define las clases, con los parámetros de concatenación regular de cadenas a partir de las entradas y las terminaciones.
  modelo v1 {raíz + terminaciones de la 1ª conjugación regular + rasgos}

Las fuentes declarativas de MPS, en cambio, están constituidas por un único tipo de archivo:

- **rcm.txt:** listado de reglas de reagrupamiento, descomposición y correspondencia, que especifica cadenas posibles de lemas con una sintaxis específica para la máquina. Puede incluir tres tipos de reglas:
  - ✓ reglas de reagrupamiento: D + N = SN

- ✓ reglas de descomposición: contracc = P + D
- ✓ correspondencia: Art = D

En este trabajo utilizamos como input de MPS el output obtenido del análisis presintáctico (morfológico) de textos en español a partir de la utilización de SMORPH. Para esta aplicación se empleó como lexicón el archivo entradas.txt elaborado por el equipo Infosur, y la correspondiente modelización desarrollada por el mismo equipo para las formas flexivas. En los parágrafos siguientes proponemos un modelo de declaración de reglas en MPS con la finalidad de desambiguar las salidas obtenidas de Smorph.

# 4. PROCESO DE DESAMBIGUACIÓN

## 4.1. Ambigüedad N / A

La lexicografía ha presentado diversos tratamientos de este tipo de ambigüedad según reconociera una o más entradas léxicas en los diccionarios para palabras tales como ‘cuerda’ (para la cual el diccionario de la R.A.E. dispone dos entradas, ‘cuerda’: N y ‘cuerdo-a’: A), o ‘capital’ (tratado como N y A en la misma entrada).

Otra cuestión se presenta cuando una misma palabra, en principio A, puede aparecer como complemento de N o como núcleo de un SN ante la ausencia del N correspondiente: es el caso de ‘buenos’ en la secuencia (1) frente a la secuencia (2)

(1) ‘los hombres buenos’
(2) ‘los buenos’

Una primera etapa, entonces, consistió en plantear la extracción de sintagmas nominales núcleos [6], a fin de poder asignar la categoría N a los sustantivos en posición de núcleos de SN, A a los adjetivos que los complementan, y ºA a los adjetivos en posición de núcleo de SN.

### *4.1.1. 1º paso: Normalización*

A fin de obtener automáticamente textos que permitiera trabajar con los SN, declaramos en un archivo rcm1.txt **reglas de descomposición** que analizaran las contracciones de modo tal que se liberara el artículo para el posterior tratamiento de los SN:

Tabla 2: Reglas de SN en archivo rcm2.txt

| Archivo rcm1.txt | | |
|---|---|---|
| %R 001% | contr1 → P + Det | ‘al’ → ‘a’ + ‘el’ |
| %R 002% | contr2 → P + Det | ‘del’ → ‘de’ + ‘el’ |

### *4.1.2. 2º paso: Declaración de reglas de SN*

El paso siguiente consistió en declarar en un archivo rcm2.txt todas las **reglas de composición** necesarias para la buena formación de SN. De esta manera, realizando una nueva ejecución de MPS con este archivo de reglas, extraeríamos todas las secuencias interpretables como SN.

Las reglas de SN quedaron declaradas de este modo:

Tabla 3: Reglas de SN en archivo rcm2.txt

| | | |
|---|---|---|
| %R 100% | D + N + A→ SN | ‘la mujer buena’ |
| %R 105% | D + N → SN | ‘la mujer’ |
| %R 110% | D + A + N → SN | ‘la buena mujer’ |

| | | |
|---|---|---|
| %R 115% | D + A +A → SN | 'la buena anciana' |
| %R 120% | D + A → SN | 'la buena' |
| %R 125% | N + A → SN | 'mujer buena' |
| %R 130% | A + N → SN | 'buena mujer' |
| %R 135% | A + A → SN | 'buena anciana' |
| %R 140% | N → SN | 'mujer' |

*4.1.3. 3º paso: Detección del núcleo en D+A+A*

Un problema aparte lo suscitó el reconocimiento del núcleo en secuencias D+A+A, ya que los adjetivos muestran diferentes comportamientos en colocación:

1.- Adjetivos que nunca son núcleo: 'buena'

'La vieja buena' frente a 'La buena vieja'
D + ºA + A D + A + ºA

y

'La buena alemana' frente a 'La alemana buena'
D + A + ºA D + ºA + A

2.- Adjetivos que son núcleo cuando aparecen en posición inicial:

'El alemán médico' frente a 'El médico alemán'
D + ºA + A D + ºA + A

3.- Adjetivos que son complemento cuando siguen a otro, pero que resultan ambiguos si lo preceden:

'El alemán joven' frente a 'El joven alemán'
D + ºA + A D + ?A + ?A

Esto nos condujo a determinar, en principio, tres clases de adjetivos:

Tabla 4: Tipos de Adjetivo

| + | **Adj1** bueno malo | **Adj2** alemán médico | **Adj3** joven viejo |
|---|---|---|---|
| **Adj1** 'bueno' 'malo' | Ø | A + ºA | A + ºA |
| **Adj2** 'alemán' 'médico' | ºA + A | ºA + A | ºA + A |
| **Adj3** 'joven' 'viejo' | ºA + A | ? + ? | Ø |

La flecha indica la dirección de lectura, el signo 'Ø' cadenas inaceptables, y el signo '?' cadenas que son ambiguas en el lenguaje natural, y por lo tanto, no puede exigirse al programa que las desambigüe.

Para esto, declaramos entonces las tres clases de adjetivos en el archivo de rasgos y en las respectivas entradas del diccionario, y luego se agregaron las reglas correspondientes en rcm2.txt (115 y 135 reemplazan a las declaradas anteriormente):

Tabla 5: Reglas de SN en archivo rcm2.txt

| | | | |
|---|---|---|---|
| %R 115% | D + Adj1 + A | → SN → D+A+ºA | ‘la buena alemana’ |
| %R 116% | D + Adj3 + Adj2 | → SN → D+?A+?A | ‘la joven alemana’ |
| %R 117% | D + Adj3 + Adj1 | → SN → D+ºA+A | ‘la joven buena’ |
| %R 118% | D + Adj2 + A | → SN → D+ºA+A | ‘la alemana joven’ |
| %R 135% | Adj1 + A | → SN → A+ºA | ‘buena alemana’ |
| %R 136% | Adj3 + Adj2 | → SN → ?A+?A | ‘joven alemana’ |
| %R 137% | Adj3 + Adj1 | → SN → ºA+A | ‘joven buena’ |
| %R 138% | Adj2 + A | → SN → ºA+A | ‘alemana joven’ |

## 4.2. Ambigüedad N / V

La etapa siguiente consistió en abordar las ambigüedades N/V, del tipo ‘amenaza’, ‘trabajo’, ‘informes’ o ‘deber’.

El problema radicaba en que las reglas declaradas para SN aplicaban la interpretación D+N en secuencias ambiguas del tipo ‘la amenaza’, ‘los informes’, ‘lo presente’, adecuada en oraciones como:

- ‘La amenaza provocó pánico’
- ‘Trajo los informes’
- ‘Lo presente es lo único que vale’

pero inadecuada en:

- ‘Juan la amenaza’
- ‘Espero que los informes’
- ‘Quiero que lo presente rápidamente’

Evidentemente, esto se complementaba con el hecho de que los pronombres ‘la’, ‘las’, ‘lo’ y ‘los’ también reciben una doble interpretación como pronombres clíticos (cl) y como artículos (art) según la secuencia que integran.

### *4.2.1. 1º paso: Desambiguación de cadenas no ambiguas en secuencia.*

El primer paso en esta nueva etapa fue evitar interpretaciones erróneas tales como:

‘lo amenaza’ → D + N

resultante de las reglas declaradas hasta aquí. Esto implicó incluir la concordancia como un dato a considerar, puesto que la falta de coincidencia de rasgos de género y número (declarados en el lexicón) nos permitió definir la interpretación en un buen número de secuencias.

Para ello, en primer término, reemplazamos las reglas 105 y 120 por las siguientes:

Tabla 6: Redefinición de reglas de SN en archivo rcm2.txt

| | | |
|---|---|---|
| %R 105% | D~art + N → SN | ‘esta mujer’ |
| %R 120% | D~art + A → SN | ‘una alemana’ |

que permitieron tratar como SN todas las secuencias de N o A precedidos por cualquier determinante a excepción del artículo (~art). Previamente se incluyeron los rasgos art y ~art (declarados en rasgos.txt como Tipos de Determinante) a cada uno de los determinantes en las respectivas entradas.

Eliminamos también la regla 140, que determinaba ‘amenaza’ como N en contextos no previstos en las demás reglas.

La cadena ambigua del tipo ‘deber’ = V Infinitivo / N resulta de tratamiento simple, puesto que aparece en contextos bien definidos:

- SN ‘el deber’, donde Art + N (para los sintagmas con det ~art aplica la regla 105).
- SP ‘sin saber’, donde P + V inf.
- SV ‘va a poder’, donde V + P + V inf (y otras frases verbales de estructura fija).

Para las frases verbales, incluimos en el archivo rcm1.txt todas las reglas de formación, de modo tal que en la primera ejecución de MPS con este archivo, no sólo quedaran analizadas las contracciones, sino también concatenadas las secuencias de SV para tenerlas disponibles en el análisis posterior. La tabla 8 muestra algunas de estas **reglas de composición**:

Tabla 7: Reglas de SV en archivo rcm1.txt

| Archivo rcm1.txt | | |
|---|---|---|
| %R 010% | mod + cj + V inf → SV | ‘tiene que estudiar’ |
| %R 020% | mod + p + V inf→ SV | ‘va a estudiar’ |
| %R 030% | mod + V inf→ SV | ‘debe estudiar’ |
| %R 040% | ser + V pp→ SV | ‘fue estudiado’ |
| %R 050% | auxtc + V pp → SV | ‘ha estudiado’ |
| %R 060% | mod + V ger → SV | ‘está estudiando’ |

Aplicamos la etiqueta ‘mod’ de manera genérica a los verbos que encabezan las frases verbales. Declaramos el rasgo ‘auxtc’ para ‘haber’ como auxiliar de tiempos compuestos. Para el verbo ‘ser’ se utiliza la etiqueta ‘ser’ en lugar de ‘v’ dado su comportamiento particular en la voz pasiva.

Incluimos también reglas más complejas incluyendo formas perfectas y pasivas de los verbos modales y auxiliares cuando ello era posible y reglas para los SV que contienen clíticos en el interior. Por ej.:

Tabla 8: Reglas de SV complejos en archivo rcm1.txt

| | | |
|---|---|---|
| %R 007% | auxtc + mod pp + cj + auxtc inf + V pp→ SV | ‘ha tenido que haberlo estudiado’ |
| %R 027% | auxtc + mod + p + auxtc inf + ser pp + V pp → SV | ‘ha debido de haber sido estudiado’ |

En total en rcm1.txt declaramos un total de 40 reglas: **38 reglas de composición** de SV y **2 reglas de descomposición** de contracciones.

Complementariamente, de esta forma quedaron desambiguadas las ocurrencias de los participios en frases verbales: ‘pasado’ (N / A / V pp), ‘vista’ ( N / V pp), etc.

Luego incluimos en rcm2.txt las reglas de secuencias no ambiguas de dos cadenas ambiguas:

Tabla 9: Reglas de secuencias no ambiguas en archivo rcm2.txt

| | | | |
|---|---|---|---|
| %R 140% | Art neut + A | → SN → D+A | ‘lo bueno’ |
| %R 141% | Art m sg + N | → SN → D+N | ‘el deber’ ‘el informe’ |
| %R 142% | Cl f + N m | → SV → Cl+V | ‘la informe’ ‘las informes’ |
| %R 143% | Cl m + N f | → SV → Cl+V | ‘lo ayuda’ ‘los ayudas’ |
| %R 144% | Cl sg + N pl | → SV → Cl+V | ‘lo ayudas’ ‘la informes’ |
| %R 145% | Cl pl + N sg | → SV → Cl+V | ‘las ayuda’ ‘los informe’ |
| %R 146% | P + V inf | → SP → P+Vinf | ‘sin saber’ ‘para saber’ |
| %R 147% | Cl + Cl + V | → SV → Cl+Cl+V | ‘se los muestra’ |
| %R 148% | Cl dat + V | → SV → Cl +V | ‘le muestra’ ‘les muestra’ |

La regla 141 permite desambiguar la ocurrencia de los infinitivos en posición de N en el SN y la 146 los que aparecen precedidos por preposición. Complementariamente, se desambigua esta ocurrencia de ‘para’ (P ‘para’ / V ‘parar’ o ‘parir’).

*4.2.2. 2º paso: Postergación de cadenas ambiguas aún en secuencia.*

Quedan como remanentes dos tipos de secuencias:

- ‘la amenaza’ ‘las amenazas’ ‘los informes’
- ‘amenaza’, ‘amenazas’. ‘informes’, ‘resumen’ sin artículo/clítico

Evidentemente, estas secuencias son desambiguables en el contexto:

- ‘Juan la amenaza.’
- ‘Juan amenaza.’

Sin embargo, pueden darse contextos que incluyan otras ambigüedades:

- ‘la amenaza causa…’
- ‘amenaza causa…’

En una secuencia como ‘La plaga causa amenazas serias para la cosecha’, 7 cadenas de 8 son ambiguas; por lo tanto, será necesario contar con análisis en etapas que extraigan sucesivamente sintagmas, dejando para etapas siguientes las interpretaciones definitivas.

Proponemos entonces una serie de reglas de postergación, que permitan suspender la asignación de una u otra interpretación hasta contar con datos del contexto. En realidad, se trata de **reglas de correspondencia**, que al pasar por cadenas ambiguas les asigna nuevamente una etiqueta de ambigüedad.

Para que MPS reconozca estas cadenas, en necesario declarar en las entradas un rasgo ‘ambNV’ (previamente incluido en rasgos.txt) a cada uno de los lemas que pueden interpretarse como N o como V [7], por ej:

```
amenaza n1/ambNV .
causa n1/ambNV .
resumen n2/ambNV .
informe n3/ambNV .
```

A continuación, declaramos las reglas de postergación (o reglas de correspondencia) en rcm2.txt:

Tabla 10: Reglas de postergación en archivo rcm2.txt

| | | | |
|---|---|---|---|
| %R 150% | Cl acus + N ambNV | → AmbSNSV | ‘la ayuda’ ‘los informes’ |
| %R 151% | N ambNV | → AmbNV | ‘ayuda’ ‘informes’ |

Una vez declaradas estas reglas, que determinan que el análisis ‘saltee’ las cadenas ambiguas no analizadas en las reglas anteriores, pueden agregarse las reglas para las cadenas que aún quedan pendientes:

Tabla 11: Reglas de SN no ambiguos en archivo rcm2.txt

| | | | |
|---|---|---|---|
| %R 160% | Art + N | → SN → D+N | ‘la página’ |
| %R 161% | Art + A | → SN → D+A | ‘la anciana’ |

Quedan cubiertas así todas las posibilidades de SN en esta primera ejecución de rcm2.txt.

*4.2.3. 3º paso: Desambiguación cadenas ambiguas aún en secuencia.*

La declaración de reglas de correspondencia respondió a dos decisiones metodológicas:

- postergar el análisis de secuencias hasta contar con datos del contexto como input,
- transformar las etiquetas que Smorph asignaba a estas secuencias en otras etiquetas, sólo visibles para una segunda aplicación de MPS.

De esta manera, en un mismo archivo, contamos con reglas que actúan sólo en la primera ejecución y otras que actúan en la ejecución siguiente. Esta metodología podría utilizarse en más ejecuciones de resultar necesario para el análisis de otras cadenas (adverbiales, por ej.).

Declaramos entonces en rcm2.txt reglas de composición, que están disponibles únicamente en la segunda ejecución de MPS con ese archivo, pues operan sobre etiquetas que arrojó MPS en el output anterior. La tabla 12 muestra algunas de esas reglas.

Tabla 12: Reglas de segunda ejecución en archivo rcm2.txt

| | | | |
|---|---|---|---|
| %R 200% | SN + AmbNV + AmbSNSV | → SN+SV+SN | 'La maestra muestra los informes' |
| %R 201% | AmbNV + AmbSNSV | → SV+SN | 'muestra los informes' |
| %R 202% | SN + AmbSNSV | → SN+SV | 'La maestra la muestra' |
| %R 210% | SV + AmbSNSV | → SN+SN | 'presentaron los informes' |
| %R 215% | SV + AmbNV | → SN+SV+SN | 'presentaron los informes' |
| %R 220% | P + AmbSNSV | → SP → P+SN | 'de los informes' |
| %R 225% | P + AmbNV | → SP → P+N | 'en resumen' |

Con estas reglas quedaron resueltos los sintagmas que contenían una (R 225), dos (R 202, 210, 215, 220) o tres cadenas ambiguas (R200 y 201).

En total fueron declaradas 40 reglas en rcm2.txt, suficientes para el tratamiento de las ambigüedades N / V que nos habíamos propuesto analizar.

Tal como está presentado, este modelo deja sin resolver ex profeso secuencias como:

- 'La amenaza.'
- 'La muestra.'
- 'Trabajo duro.'

que serán consideradas AmbSNSV, es decir, secuencias que pueden tanto ser sintagmas nominales como verbales de acuerdo con el contexto en que aparezcan.

En realidad, este era uno de nuestros objetivos, puesto que en los modelos estadísticos se les asigna siempre la interpretación D+N, aún cuando es perfectamente posible encontrar contextos en los que la interpretación Cl+V sea la adecuada:

"Juan no soporta a su mujer. La agrede. La amenaza. Pero cuando ella se aleja, la extraña. La asfixia. La llama."

Como puede apreciarse, 'la amenaza', 'la asfixia' y 'la llama' dependen del contexto extraoracional para desambigüarse, y está fuera del alcance de un análisis oracional. Incluso en el análisis de títulos, la interpretación de estas secuencias dependerá de factores pragmáticos que también están más allá del ámbito de la oración.

## 4. EVALUACIÓN DEL MODELO

Se evaluó este modelo en un archivo de 10200 palabras. Sobre 422 palabras ambiguas, 380 resultaron desambiguadas con las reglas declaradas. Las 42 palabras restantes corresponden a ambigüedades no consideradas aún. De las 380 desambiguadas, sólo se detectó un error, en una

secuencia ‘sustantivo’ ‘guión’ ‘sustantivo’ (‘ecuación esfuerzo-premio’) no considerada en el modelo. En síntesis, la evaluación del modelo da como resultado:

Precisión: 99,7%
Cobertura: 89,8%

Las tablas 12, 13, 14 y 15 muestran algunos fragmentos de los archivos smorph.txt y smorph_g.txt, obtenidos como salida tras cada ejecución de Smorph y MPS:

Tabla 13: Fragmentos de smorph.txt después de ejecutar Smorph.

```
…
'ante'.
[ 'ante', 'EMS','prep'].
'posibles'.
[ 'posible', 'EMS','adj', 'GEN','_', 'NUM','pl'].
'choques'.
[ 'choque', 'EMS','nom', 'GEN','masc', 'NUM','pl', 'TAMB','ambNV'].
[ 'chocar', 'EMS','v', 'MODOV','subj', 'PERS','2a', 'NUM','sg', 'TPO','pres', 'TR','irr', 'TC','c1'].
…
'debió'.
[ 'deber', 'EMS','v', 'MODOV','ind', 'PERS','3a', 'NUM','sg', 'TPO','prets', 'TR','r', 'TC','c2'].
'ser'.
[ 'ser', 'EMS','nom', 'GEN','masc', 'NUM','sg', 'TAMB','ambNV'].
[ 'ser', 'EMS','ser', 'MODOV','infin', 'TR','irr', 'TC','c2'].
'internado'.
[ 'internar', 'EMS','v', 'MODOV','part', 'GEN','masc', 'NUM','sg', 'TR','r', 'TC','c1'].
…
'Existían'.
[ 'existir', 'EMS','v', 'MODOV','ind', 'PERS','3a', 'NUM','pl', 'TPO','imp', 'TR','r', 'TC','c3'].
'informes'.
[ 'informe', 'EMS','nom', 'GEN','masc', 'NUM','pl', 'TAMB','ambNV'].
[ 'informar', 'EMS','v', 'MODOV','subj', 'PERS','2a', 'NUM','sg', 'TPO','pres', 'TR','r', 'TC','c1'].
'que'.
[ 'que', 'EMS','rel'].
[ 'que', 'EMS','sub'].
'advierten'.
[ 'advertir', 'EMS','v', 'MODOV','ind', 'PERS','3a', 'NUM','pl', 'TPO','pres', 'TR','irr', 'TC','c2'].
…
```

Tabla 14: Fragmentos de smorph_g.txt después de aplicar MPS con rcm1.txt.

```
'ante'.
[ 'ante', 'EMS','prep'].
'posibles'.
[ 'posible', 'EMS','adj', 'GEN','_', 'NUM','pl'].
'choques'.
[ 'choque', 'EMS','nom', 'GEN','masc', 'NUM','pl', 'TAMB','ambNV'].
[ 'chocar', 'EMS','v', 'MODOV','subj', 'PERS','2a', 'NUM','sg', 'TPO','pres', 'TR','irr', 'TC','c1'].
…
'debió ser internado'.
[ 'deber ser internar', 'EMS', 'SV', 'EMS', 'V+Vser+VPpio' ].
…
'Existían'.
[ 'existir', 'EMS','v', 'MODOV','ind', 'PERS','3a', 'NUM','pl', 'TPO','imp', 'TR','r', 'TC','c3'].
'informes'.
[ 'informe', 'EMS','nom', 'GEN','masc', 'NUM','pl', 'TAMB','ambNV'].
[ 'informar', 'EMS','v', 'MODOV','subj', 'PERS','2a', 'NUM','sg', 'TPO','pres', 'TR','r', 'TC','c1'].
'que'.
```

```
[ 'que', 'EMS','rel'].
[ 'que', 'EMS','sub'].
'advierten'.
…
```

Tabla 15: Fragmentos de smorph_g.txt después de aplicar MPS con rcm2.txt.

```
…
'ante'.
[ 'ante', 'EMS', 'prep' ].
'posibles choques'.
[ 'posible choque', 'EMS', 'SN', 'EMS', 'A+ºN' ].
…
'debió ser internado'.
[ 'deber ser internar', 'EMS', 'SV', 'EMS', 'V+Vser+VPpio' ].
…
'Existían'.
[ 'existir', 'EMS', 'SV', 'EMS', 'V' ].
'informes'.
[ 'informe', 'EMS', 'AmbNV' ].
'que'.
[ 'que', 'EMS', 'rel' ].
[ 'que', 'EMS', 'sub' ].
'advierten'.
[ 'advertir', 'EMS', 'SV', 'EMS', 'V' ].
…
```

Tabla 16: Fragmentos de smorph_g.txt después de ejecutar nuevamente MPS con rcm2.txt.

```
…
'ante posibles choques'.
[ 'ante posible choque', 'EMS', 'SP', 'EMS', 'P+SN' ].
…
'debió ser internado'.
[ 'deber ser internar', 'EMS', 'SV', 'EMS', 'V+Vser+VPpio' ].
…
'Existían informes'.
[ 'existir informe', 'EMS', 'SV+SN', 'EMS', 'ºV+ºN' ].
'que'.
[ 'que', 'EMS', 'rel' ].
[ 'que', 'EMS', 'sub' ].
'advierten'.
[ 'advertir', 'EMS', 'SV', 'EMS', 'V' ]
…
```

# 5. CONCLUSIONES Y PROYECCIONES

Entre los logros de esta modelización contamos la complementación que se produce entre las reglas declaradas, de manera tal que logramos desambiguar las secuencias elegidas, pero a la vez se producen otras desambiguaciones, lo que va a cotando progresivamente las ambigüedades remanentes. Esto es lo que ocurrió con las cadenas Cl/Art, y con los verbos en infinitivo. Lo mismo sucede con las cadenas ‘para’, ya desambiguada como preposición cuando aparece seguida por V inf, o con ‘una’ (Det / V) interpretado como Det en SN.

Los cálculos de precisión y cobertura resultan superiores a los valores arrojados en las evaluaciones que hemos hecho de otras herramientas informáticas.

En próximas etapas abordaremos otras ambigüedades referidas a preposiciones, adverbios, conjunciones, etc.

La integración de esta modelización a herramientas más complejas de corrección ortográfica, gramatical, e incluso en análisis de textos de aprendices de español como lengua materna o segunda lengua parece viable a mediano o largo plazo.

## Referencias

[1] Utilizamos el término “cadena” para referirnos a la sucesión de caracteres que formas un lexema o palabra reconocible del español: ‘casa’. Con “secuencia” nos referimos a una sucesión de “cadenas” o palabras: ‘la casa rosada’.

[2] Es imposible abordar aquí la discusión acerca de clases de palabras. Por esta razón, para la asignación de las etiquetas correspondientes a cada categoría gramatical, tomamos como referencia el *Diccionario de la Real Academia Española*, 22ª Edición, http://www.rae.es/rae.html.

[3] N = Nombre o sustantivo, V = Verbo, V inf = Verbo infinitivo, V pp = Verbo participio, A = Adjetivo, Adv = Adverbio, P = Preposición, D = Determinante, Pr = Pronombre, Cj = Conjunción, Intj = Interjección.

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

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

[6] Para un tratamiento más detallado de los sintagmas núcleos, Bès G., Lamadon L. y Trouilleux F. “Verbal chunks extraction in French using limited resources”. *arXiv:cs.CL/0408060 v1*, 2004.

[7] En esta lógica seguimos a Solana Z. y Bès G. “Extracción del sintagma verbal núcleo y resolución de ambigüedades en la asignación categorial”. *Revista de Letras* Nº 9, Vol. de Estudios Lingüísticos. UNR. Rosario, 2004.