

# AMBILIC. EL DESAMBIGUADOR LINGÜÍSTICO DEL CORPUS DEL IULA (UPF)

LLUÍS DE YZAGUIRRE MAURA, CARMÉ BACH MARTORELL,  
ANNA MATAMALA RIPOLL, NÚRIA CASTILLO IGEA  
Y EUGÈNIA USTRELL PEÑAFIEL  
*Universitat Pompeu Fabra*

## INTRODUCCIÓN

AMBILIC es un programa de desambiguación de base lexicomorfológica para el castellano y para el catalán que se utiliza en la cadena de tratamiento del corpus técnico del Institut Universitari de Lingüística Aplicada (IULA) de la Universidad Pompeu Fabra (Bach *et al.* 1997). Se aplica a ficheros que han sido analizados morfológicamente con el programa PALIC (De Yzaguirre 2000a), con la finalidad de eliminar el máximo número de ambigüedades con reglas únicamente lingüísticas. Los resultados demuestran que AMBILIC reduce las ambigüedades de un 170% —porcentaje inicial, similar, v.g., al del francés (Tzoukermann y Radev 1996)— a un 110% aproximadamente. Las que quedan después de aplicar AMBILIC se resuelven mediante un desambiguador estocástico.

En esta comunicación pretendemos presentar de modo general el funcionamiento interno del programa de desambiguación AMBILIC (segundo apartado) y los tipos de reglas que utiliza (tercer apartado), así como posibles mejoras que nos planteamos de cara al futuro (cuarto apartado).

## FUNCIONAMIENTO DE AMBILIC Y RESULTADOS

AMBILIC se puede aplicar tanto a documentos del corpus del IULA como a documentos de otras fuentes. En el primer caso, los textos están marcados estructuralmente (Vivaldi *et al.* 1996) y han pasado por un preproceso en el que se han detectado y etiquetado algunas unidades particulares, como abreviaturas, números, locuciones, entidades nominales, etcétera; han sido analizados morfológicamente con PALIC y contienen el/los lema(s) y la(s) categoría(s) de cada forma. En el segundo caso, cuando los textos a analizar no son del corpus del IULA, se incorpora a los documentos un marcaje estructural mínimo y se procesan con el programa PALIC para obtener los lemas y las categorías posibles de cada forma.

Las categorías propuestas para las distintas palabras en ambos casos están codificadas mediante unas etiquetas establecidas para el corpus del IULA siguiendo las directrices del proyecto EAGLES (Morel *et al.* 1997).

En cuanto al etiquetario, conviene puntualizar que, en el marco del proyecto Corpus, se distingue entre ambigüedad y subespecificación. Si una palabra tiene dos interpretaciones, pero ambas corresponden al mismo lema, se le atribuye una etiqueta más genérica, llamada subespecificada. Así, el sustantivo “viernes” es etiquetado N5M6, que significa “nombre común masculino de género subespecificado”, o sea indistintamente singular o plural. Lo mismo ocurre con formas verbales como “lleva” o “diría”. Sólo se consideran ambiguas y sólo se pretende desambiguar formas con varias interpretaciones atribuibles a lemas distintos.

AMBILIC resuelve las ambigüedades utilizando gramáticas locales (Silberztein 1993) en forma de bancos de reglas del catalán o del castellano. Debido a su posición en la cadena de procesamiento del corpus del IULA, las reglas están formuladas restrictivamente para que no sobreactúen en ningún caso. Después de aplicar el programa a los textos, el resultado es un documento donde están desactivados los lemas y las categorías que no son pertinentes según el contexto. El programa lo discrimina gráficamente indicando con minúscula la etiqueta que ha desactivado, mientras el resto de opciones siguen representadas con mayúscula.

En lo que concierne a los resultados, tal como se ha señalado, el programa reduce las ambigüedades de un 170% a un 110%. En un estudio preliminar con un corpus de cuarenta mil palabras, el margen de error de AMBILIC y el desambiguador estocástico combinados fue de un 3%, un porcentaje dentro de los estándares habituales. El programa se ha aplicado a un total de más de cinco millones de palabras con un funcionamiento satisfactorio.

## TIPOS DE REGLAS

AMBILIC opera con reglas que han sido formuladas restrictivamente para que no sobreactúen en ningún caso de acuerdo con la posición que ocupa en la cadena de procesamiento del Corpus del IULA; siempre es preferible que no actúen si el contexto no está muy bien definido y que las ambigüedades se resuelvan en el paso siguiente, la desambiguación por métodos estocásticos.

Existe una delimitación previa de los rasgos aplicables a las categorías mayores. Así, AMBILIC asocia a cada rasgo de las etiquetas morfológicas un conjunto de descriptores de la forma siguiente:

Tabla 1. Descriptores de rasgos

VC {MF} {SP} /categoría, modo, género, número
JQ{MF6} {SP6} /categoría, clase, género, número

En el primer ejemplo se indica que la categoría “V” del modo “C” (que corresponde a un verbo en modo participio) puede tener género masculino o femenino y que el número puede ser singular o plural. En el segundo, la categoría “J” de la clase “Q” (que corresponde a un adjetivo calificativo) puede tener género masculino, femenino o pendiente de especificar y el número puede ser singular, plural o pendiente de una especificación posterior. Estos descriptores, basados en el etiquetario del IULA, son los que permiten redactar reglas, así como formular condiciones basadas en formas, lemas, número de lemas, posición en el contexto (inicial, final...).

Las reglas de desambiguación pueden tomar distintas decisiones al mismo tiempo y constan de dos partes: la condicional y la ejecutiva. La primera parte filtra si la regla se tiene que ejecutar o no, y la segunda sirve para eliminar un lema o unos lemas o para eliminarlos todos menos uno.

Para leer las reglas de desambiguación, se debe tener en cuenta que, en cada línea, la cifra indica la distancia de la palabra a la que se refiere la condición o ejecución respecto a la primera del contexto que se está estudiando. A continuación, se sitúa el descriptor de la variable lingüística (en el ejemplo de la Tabla 2, “n\_lemas” o “categoría”) seguido de un delimitador, que puede ser “=” (es) o “#” (no es). Al final de la línea aparece el valor que

**Tabla 2.** Parte condicional y parte ejecutiva

<pre>* Regla 4005 -&gt; "el mejor de" 0,categoría=A 1,n_lemas=2 1,categoría=J 1,categoría=D 2,categoría=P \ 1,categoría=J /</pre>	<p>Interpretación:</p> <p>Si una palabra que es determinante va seguida de una palabra con dos lemas (adjetivo y adverbio), seguida a su vez de una preposición, entonces la palabra siguiente al determinante es adjetivo.</p>
---	---

tiene que presentar la variable para actuar, codificado mediante unas etiquetas. La contrabarra "\ " indica que se acaba la parte condicional y empieza la ejecutiva, y la barra "/" indica el final de la regla de desambiguación.

AMBILIC opera con paquetes de reglas que se agrupan cuando tienen rasgos similares: esta estrategia facilita la ejecución del programa y también permite al usuario activar o desactivar un determinado paquete de reglas teniendo en cuenta las características textuales del documento que está analizando. Hay distintos criterios de agrupación de las reglas de desambiguación: por un lado, se pueden reunir en un mismo paquete según sean de tipo gramatical, de tipo léxico o híbridas. Por otro lado, se pueden agrupar las reglas en base a características dialectales (reglas dialectales) o temáticas (reglas temáticas). Ejemplos de este último tipo de reglas serían las referentes a los nombres de notas o a los nombres de letras.

Los paquetes de reglas del programa AMBILIC se aplican a todas las unidades que se encuentran entre dos signos de puntuación siguiendo un orden secuencial, de izquierda a derecha. Si cuando se llega a la posición final se ha modificado alguna palabra, se vuelven a aplicar todas las reglas del paquete y así sucesivamente hasta que no se haya producido ninguna modificación. En otras palabras, la condición de finalización del ciclo es el ciclo vacío. Por otro lado, una de las características del programa es que el usuario puede determinar qué paquetes de reglas quiere aplicar, en qué orden e incluso cuáles se van a aplicar más de una vez.

Las reglas de desambiguación pueden ser de dos tipos: acontextuales o contextuales. En los dos subapartados siguientes se explican más detalladamente y se ofrecen ejemplos.

#### Reglas acontextuales

Las reglas acontextuales son las que eliminan alguna interpretación sin tener en cuenta el contexto. Presentamos un ejemplo en la Tabla 3:

**Tabla 3.** Regla acontextual

<pre>* Regla 0001 -&gt; "a" 0,lema=A \ 0,categoría=P /</pre>
--

En la regla de la Tabla 3 se indica que, cuando se encuentra un lema "a", se desambigua como preposición. De este modo se anula la posibilidad de que "a" se desambigüe como sustantivo.

Las reglas acontextuales se utilizan en casos como los nombres de notas o los nombres de letras, en vez de eliminar estas unidades del diccionario del analizador morfológico. Tanto si utilizamos reglas acontextuales como si eliminamos determinadas unidades del diccionario el resultado es el mismo, pero este sistema permite que, en unos textos concretos, el usuario pueda desactivar determinadas reglas. Por ejemplo, podemos formular unas reglas acontextuales según las cuales las unidades "a", "de", "e", "ca", "ele", "o", "erre", "ese" o "te" no sean sustantivos, anulando de este modo la posibilidad de que correspondan a nombres de letras, pero el usuario tiene la opción de desactivar este paquete de reglas cuando desambigüe un texto de temática lingüística, ya que entonces es probable que estas unidades aparezcan como nombres de letras.

#### Reglas contextuales

El desambiguador AMBILIC también trabaja con reglas contextuales que, tal como su propio nombre indica, tienen en cuenta el contexto al formular las condiciones. Presentamos varios ejemplos en la tabla siguiente:

**Tabla 4.** Reglas contextuales

<pre>*Regla 4019 -&gt; "la requerida" 0,n_lemas=2 0,categoría=A 0,categoría=R 1,n_lemas=1 1,modo=C \ 0,categoría=A /</pre>	<p>Interpretación:</p> <p>Si una palabra tiene dos lemas (uno de ellos determinante y el otro pronombre) y la palabra siguiente tiene un solo lema y es participio, entonces la primera palabra se desambigua como determinante.</p>
<pre>* Regla 4026 -&gt; "cuyos intereses" 0,lema=cuyo 1,n_lemas=2 1,categoría=V 1,categoría=N \ 1,categoría=N /</pre>	<p>Interpretación:</p> <p>Si una palabra con lema "cuyo" va seguida de una palabra con dos lemas (verbo y nombre), entonces esta última palabra se desambigua como nombre.</p>
<pre>* Regla 4033-&gt; pron.#det.? más no verbo 0,n_lemas=2 0,categoría=R 0,categoría=A 1,categoría#V \ 0,categoría=A /</pre>	<p>Interpretación:</p> <p>Si una palabra con dos interpretaciones, una de ellas pronombre y la otra determinante, va seguida de una palabra que no es verbo, entonces esa palabra es solamente determinante.</p>

El formalismo de las reglas de AMBILIC permite expresar variables coincidentes (como “X” en las reglas 4030 y 4043 que presentamos en la Tabla 5) o que una determinada variable puede tener cualquier valor (como “tiempo=W”, en la regla 4004, que indica que se trata de una forma verbal conjugada, lo que excluye las formas nominales de los verbos y el resto de categorías). También permite expresar una lista de lemas o de formas en una única condición; por ejemplo, ante coincidencias concurrentes del tipo “derechos humanos, individuales, colectivos, inenajenables, fundamentales” se puede proponer una regla como la que presentamos en la Tabla 5.

**Tabla 5.** Variables coincidentes

<p>* Regla 4030 -&gt; determinante seguido de nombre, adjetivo o especificador</p> <p>0,categoría=A 1,n_lemas=1 1,categoría=X{NJE} \ 0,categoría=A /</p>	<p>Interpretación:</p> <p>Si una palabra que pueda ser determinante va seguida de una palabra unívoca cuya categoría es nombre, adjetivo o especificador, entonces esa palabra es solamente determinante.</p>
<p>* Regla 4043 “el aceptante”</p> <p>0,categoría=A 1,n_lemas=2 1,categoría=X{NJ} 1,modo=G \ 0,categoría=A 1,categoría=X /</p>	<p>Interpretación:</p> <p>Si un determinante va seguido de una palabra con dos lemas, uno de ellos nombre o adjetivo y el otro gerundio, entonces el determinante sólo es determinante y la siguiente palabra sólo es no gerundio.</p>
<p>* Regla 4004 -&gt; “representan apenas”</p> <p>0,tiempo=W 0,n_lemas=1 1,mot=apenas \ 1,categoría=D /</p>	<p>Interpretación:</p> <p>Si una forma verbal personal con un solo lema va seguida de la palabra “apenas”, entonces “apenas” es adverbio.</p>
<p>0,lema=derecho 0,género=X {SP} 1,lema={humano, individual, colectivo, inenajenable, fundamental} 1,género=X \ 0,categoría=N 1,categoría=J /</p>	<p>Interpretación:</p> <p>Si estas dos palabras son consecutivas y concuerdan en género, serán consideradas respectivamente nombre y adjetivo.</p>

### Casos especiales

En este apartado nos fijaremos en algunos casos especiales, como las reglas manuales, las ambigüedades segmentales, el tratamiento de las palabras con mayúscula y el tratamiento de algunas ambigüedades entre personas verbales y sustantivos.

En ocasiones, interesa trabajar con textos desambiguados totalmente. Para conseguirlo, no se procesan estos textos con el desambiguador estocástico, sino que se elaboran una *reglas manuales* que se aplican después de la lematización y la desambiguación lingüística. A partir de esta desambiguación manual, AMBILIC puede generar bancos de reglas que sirven para reprocesar los documentos. Este conjunto de reglas manuales forman un paquete especial que se puede aplicar después de las reglas automáticas.

En cuanto a las *ambigüedades segmentales*, el programa permite que se formulen reglas específicas para resolverlas (De Yzaguirre 2000b). Nos referimos a ambigüedades como las que presenta la forma verbal “verse” (que se puede interpretar como forma del verbo “versar” o como forma del verbo “ver” más enclítico), llamadas segmentales porque plantean al módulo de *text-handling* el problema de cómo segmentarlas.

Además de las reglas manuales y las ambigüedades segmentales, el programa contempla dos casos especiales que no están formalizados como reglas sino que están incorporados como procesos. Sin embargo, el programa los reconoce como paquetes de reglas y es el usuario quien escoge si tienen que actuar o no. Nos referimos a los procesos que afectan al tratamiento de las palabras con mayúscula y a las reglas que tratan las ambigüedades entre nombre y formas verbales en primera y segunda persona del singular.

- (a) *Las mayúsculas*: si se detecta una palabra con mayúscula y una de las interpretaciones es que corresponde a un sustantivo, entonces el programa la desambiguará como sustantivo. Nos referimos a unidades como “Juramento Hipocrático”, en la que “juramento” puede ser sustantivo o primera persona del presente de indicativo del verbo “juramentar”. Si se activa el procedimiento de desambiguación por mayúsculas, el programa lo desambigua automáticamente como sustantivo. Otro ejemplo se encuentra en la frase “Me gustan lugares como La Puebla”. Una vez más, el programa desambigua “Puebla” como sustantivo y no como tercera persona del presente de indicativo de “poblar” porque detecta las mayúsculas.
- (b) *Las personas verbales*: a menudo sucede que, en un texto técnico redactado íntegramente en tercera persona, ciertas ambigüedades tienen entre sus interpretaciones una forma verbal en primera o segunda persona. El procedimiento de desambiguación de personas verbales, en caso de ser activado por el usuario, analiza si en la totalidad del documento hay primeras o segundas personas verbales no ambiguas, en cuyo caso se inhibe de actuar; por el contrario, si las únicas primeras o segundas personas halladas corresponden siempre a formas ambiguas, como “juramento”, “base”, “cierre” o “suplemento”, se elimina la interpretación verbal, siempre y cuando el documento tenga una cierta extensión y una cierta proporción de formas verbales no ambiguas. Este procedimiento actúa siempre que el texto contenga cincuenta formas verbales; si es así, dos terceras partes del total de los verbos tienen que ser unívocas en esta etapa del análisis.

Es evidente que en un programa de estas características hay muchos aspectos que podrían mejorarse. Presentamos cinco de las mejoras que consideramos más relevantes, pendientes de desarrollo.

- (a) *Desambiguación de base sintáctica*: se trata de interponer, entre el desambiguador morfológico y el estocástico, un *parser* sintáctico parcial o *chunker* (Hindle 1994), al cual se le pasen tantas copias de una misma frase como combinaciones se puedan formar a partir de las ambigüedades aún presentes. El *chunker* nos devolverá resultados más o menos satisfactorios para cada una de las interpretaciones. Conservando los mejores resultados, una parte de las ambigüedades pueden ser eliminadas por cuanto todas las soluciones preferidas comparten la misma interpretación de una determinada ambigüedad. De momento no es posible aplicar esta técnica porque no disponemos de un *chunker* suficientemente veloz para analizar los miles de combinaciones producidas por cada frase del corpus técnico del IULA. Sin embargo, dentro de poco tiempo las mejoras en los equipos usados y en los *parsers* harán rentable esta técnica.
- (b) *Desambiguación de base temática*: se trata de implementar un procedimiento que permita detectar automáticamente si el tema de un documento desaconseja aplicar un paquete de reglas concreto. Nos referimos, por ejemplo, a los ya citados de notas musicales o nombres de letras.
- (c) *Mejoras de base dialectal*: el programa tendría que poder detectar automáticamente aquellas variantes dialectales de un texto para las que existan paquetes de reglas. Así, por ejemplo, en un texto escrito en el Perú, “lustrada” tendría que figurar como participio del verbo “lustrar” y como sustantivo. En cambio, en un texto escrito en otra variedad, se podría desambiguar automáticamente como participio.
- (d) *Mejoras de base terminológica*: para poder explotar a fondo la desambiguación de base temática, habría que desarrollar un programa que fuera capaz de reutilizar terminografías —especialmente en los términos sintagmáticos— para generar reglas de desambiguación para un determinado ámbito terminológico. Por ejemplo, si se está desambiguando un texto sobre informática y se localiza la unidad “base de datos”, se tendría que poder desambiguar “base” como sustantivo y no como forma verbal.
- (e) *Resolución de la subespecificación*: como hemos visto, las reglas sólo eliminan interpretaciones de entre las atribuidas por el lematizador, pero no añaden ni modifican nada. Así, pues, queda por resolver la cuestión de las subespecificaciones. Nos referimos a casos como la forma verbal “andaba”, en la que la persona es primera o tercera. Las condiciones contextuales de las reglas de desambiguación pueden servir de manera similar para la resolución de la subespecificación, con la diferencia que su parte ejecutiva debe cambiar un rasgo genérico por otro de explícito.

#### CONCLUSIONES

A modo de conclusión, destacamos los puntos más importantes de esta comunicación:

- a) AMBILIC es un desambiguador lingüístico para el catalán y el castellano que, a partir de reglas locales formuladas restrictivamente, reduce las ambigüedades de un 170% a un 110%.

- b) AMBILIC se puede aplicar tanto a documentos del corpus del IULA como de otras fuentes, siempre que hayan sido procesados con PALIC.
- c) Los bancos de reglas se ordenan por paquetes que permiten aplicarlas recursivamente.

#### BIBLIOGRAFÍA

- Armstrong, S.; Robert, G. y P. Bouillon. 1998. “Building a language model for POS tagging”, [Documento de Internet, disponible día 10.06.00, en <http://issco-www.unige.ch/tools>].
- Bach, C.; Saurí, R.; Vivaldi, J. y M.T. Cabré. 1997. *El Corpus de l’IULA: descripció*. Papers de l’IULA. Sèrie informes, 17. Barcelona: IULA, Universitat Pompeu Fabra.
- Badia, T.; Pujol, M.; Tuells, T.; Vivaldi, J.; De Yzaguirre, L. y M.T. Cabré. 1998. “IULA’s LSP Multilingual Corpus: Compilation and Processing”. Comunicación presentada en la conferencia ELRA. Granada, mayo de 1998. [Documento de Internet, disponible día 10.06.00, en <http://www.iula.upf.es/corpus/corpubca.htm>].
- De Yzaguirre, L.; Matamala, A. y M.T. Cabré. 2000a. “El lematizador PALIC del IULA (UPF)”. Comunicación presentada en el XVIII Congreso AESLA, Barcelona, mayo 2000.
- De Yzaguirre, L.; Torner, S. y A. Matamala. 2000b. “El tratamiento automático de las ambigüedades segmentales del castellano”. Comunicación presentada en el XVIII Congreso AESLA, Barcelona, mayo 2000.
- Hindle, D. 1994. “A Parser for Text Corpora”. *Computational Approaches to de Lexicon*. Eds B.T.S. Atkins y A. Zampolli. Oxford: Oxford University Press.
- Morel, J.; Torner, S.; Vivaldi, J.; De Yzaguirre, L. y M.T. Cabré. 1997. *El corpus de l’IULA: etiquetaris*. Papers de l’IULA. Sèrie Informes, 18. Barcelona: IULA, Universitat Pompeu Fabra.
- Silbersztein, M. 1996. *Dictionnaires électroniques et analyse automatique de textes: le système INTEX*, París: Masson.
- Tzoukermann, É. y D. R. Radev. 1996. “Using word class for part-of-speech disambiguation”. Actas del Fourth Workshop on Very Large Corpora. Eds. E. Ejerhed y I. Dagan. Copenhagen.
- Vivaldi, J.; De Yzaguirre, L.; Solé, X. y M.T. Cabré. 1996. *Marcatge estructural i morfosintàctic del Corpus Tècnic amb l’estàndard SGML*. Papers de l’IULA. Sèrie Informes, 1. Barcelona: IULA, Universitat Pompeu Fabra.