

Sistema de extracción de candidatos a término
YATE:
manual de utilización

Jorge Vivaldi Palatresi

Papers de l'IULA. Sèrie Informes, 43

Barcelona

Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada

2003

BIBLIOTECA DE CATALUNYA. DATOS CIP



Dirección de las Publicaciones del IULA: Mercè Lorente Casafont

Coordinación de las Publicaciones del IULA: Lluís Codina, Gemma Martínez

Edición preliminar: Mayo de 2003

© el autor

© Institut Universitari de Lingüística Aplicada

La Rambla, 30-32

08002 Barcelona

**Sistema de extracción de candidatos a término YATE:
manual de utilización**

Jorge Vivaldi Palatresi
jorge.vivaldi@upf.edu

Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra
Barcelona

Índice

1	¿Qué es <i>YATE</i> ?	1
2	Funcionamiento básico	1
3	Requisitos para su utilización	2
3.1	Requisitos software generales	2
3.2	Requisitos específicos del programa	3
3.3	Requisitos específicos para los textos a procesar	3
4	Ejecución	3
4.1	Procedimiento básico	4
4.2	Opciones del programa <code>yate.pl</code>	5
4.2.1	Documento a procesar	6
4.2.2	Resultados obtenidos	6
4.2.3	Configuración del módulo de selección de CAT	6
4.2.4	Adaptación de <i>YATE</i> a diferentes dominios	10
4.3	Opciones del programa <code>yateol.pl</code>	14
4.3.1	Documento a procesar	15
4.3.2	Patrón a procesar	15
4.3.3	Cálculo de precisión y cobertura	15
4.4	Visualización de los resultados	16
4.5	Limitaciones	17
5	Procedimientos auxiliares	18
5.1	Obtención del fichero índice	18
5.2	Datos necesarios para el cálculo de precisión y cobertura	18
6	Referencias	19

1 ¿Qué es YATE?

YATE¹ es un sistema de extracción de candidatos a término nominales (CAT) en textos de medicina que han sido procesados previamente con las herramientas del *Corpus Tècnic* (CT) del IULA. Sus características más relevantes son:

- ➔ la obtención de una lista ordenada de candidatos a términos mediante la combinación de los resultados obtenidos con diferentes métodos de extracción y
- ➔ la utilización de información semántica en los procesos de extracción.

2 Funcionamiento básico

Una de las características de YATE es la combinación de estrategias heterogéneas para la extracción de la terminología incluida en textos especializados. Esto significa que varios módulos analizan los CAT mediante el uso de técnicas diferentes. Posteriormente, la propuesta de cada uno de ellos se combina para formular un resultado único.

Otra de las particularidades de YATE es la utilización de información semántica. El uso de este tipo de datos está plenamente justificado a partir de la definición misma de unidad terminológica. Es decir, la necesidad de comprobar que un cierto candidato pertenece o no al dominio de interés del usuario final.

Estas ideas se concretan en la arquitectura mostrada en la Figura 1. Esta figura muestra cómo, después de analizar lingüísticamente el texto a explorar y después de obtener la lista completa de los CAT, actúan varios módulos (análisis de contexto, formantes cultos, etc.) que analizan los CAT mediante el uso de técnicas heterogéneas. Posteriormente, existe un módulo específico que se encarga de combinar las propuestas de todos los módulos de análisis.

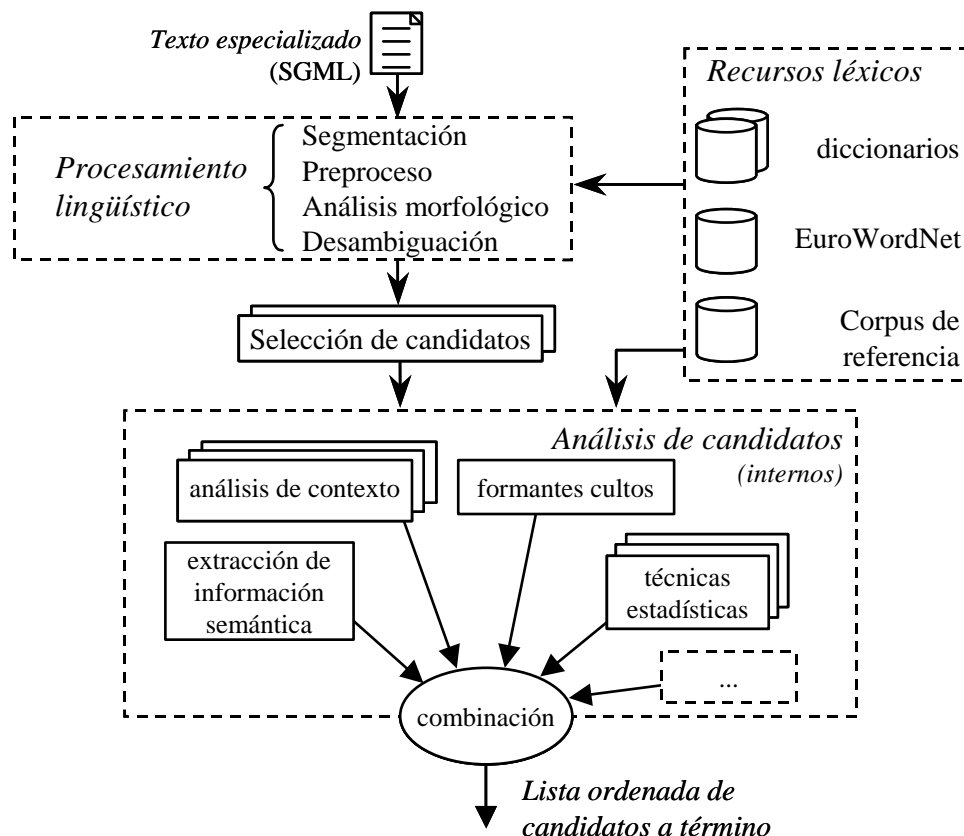
La información semántica se obtiene mediante la utilización de la base de datos *EuroWordNet*² (EWN). Aunque ésta incluya principalmente el vocabulario propio de la lengua general tiene un número de entradas sorprendentemente alto del dominio médico y otros dominios. De todas maneras, existen mecanismos que permiten ampliar esta cobertura para dar un soporte adecuado a la metodología propuesta.

El resultado de YATE será, para cada uno de los patrones considerados, una lista de CAT ordenada en función de su *termhood* o grado en que una unidad léxica representa un término pertinente al dominio considerado.

¹ Yet Another Term Extractor.

² EWN [Vossen P., 1999] es una base de datos léxica multilingüe de propósito general basada en el WN de Princeton [Fellbaum, 1998] que abarca tanto el español y el catalán como otras lenguas europeas. Cada idioma tiene su propia estructura *WordNet*, mientras que todas las lenguas incluidas en el proyecto están enlazadas entre sí por medio de una estructura común. Un *WordNet* está estructurado en unidades léxico-semánticas, o *synsets*, ligadas entre sí mediante relaciones semánticas básicas. Un *synset* es un conjunto de palabras que funcionan como sinónimos y que pueden intercambiarse en ciertos contextos. Sinonimia e hiperonimia son dos relaciones básicas tanto de WN como de EWN. La primera se utiliza en la creación de los *synsets*, ya comentada, mientras que la segunda es una relación básica empleada para vincular distintos *synsets* entre sí. WN y EWN sólo incluyen palabras que pertenezcan a algunas de las categorías siguientes: nombre, adjetivo, verbo y adverbio (YATE sólo utiliza las dos primeras).

Figura 1. Arquitectura de *VATE*.



3 Requisitos para su utilización

3.1 Requisitos software generales

Para poder ejecutar el conjunto programas que forman *VATE* es necesario disponer cómo mínimo de los siguientes recursos de software:

- Intérprete Perl con los siguientes módulos instalados:
 - DBI
 - DBD-Mysql
 - Storable
 - HTML-Entities
 - GD y GDGraph (útiles sólo si se tienen los datos necesarios para calcular precisión y cobertura)
- Conexión telemática con el ordenador “rc17.upf.es”.
- Dar valor a la variable de entorno SGML_SEARCH_PATH. Esta variable debe incluir la cadena “c:\tmp”.

El sistema ha sido desarrollado y probado en los entornos Linux (distribución RedHat 7.3) y Windows2000, aunque probablemente funcionaría en cualquier otro que cumpliera con los requisitos antes mencionados. La versión del intérprete Perl utilizada en las pruebas es la ActiveState 5.6.1.

3.2 Requisitos específicos del programa

Para que *YATE* pueda funcionar necesita también acceder a los siguientes ficheros auxiliares:

Contenido	Nombre por defecto
Fronteras de dominio	DomainMedic
Propiedades relevantes al dominio	DomainMedicAdj
Características semánticas que hacen que un CAT con patrón nombre-adjetivo sea considerado cómo válido.	DomainMedicCombineNA
Lista ordenada de formantes	FormanteO.txt
Lista de prefijos	Prefix.txt
Lista de exclusión (stop list) para los candidatos a término con el patrón NPN	Stoplist_nsp.txt

El sistema tiene ya previsto el acceso a una versión por defecto de estos ficheros. De todas maneras, en ciertas condiciones, el usuario puede construir sus propios ficheros con información personalizada. Un ejemplo de esta situación es cuando el usuario quiere definir su propio dominio de aplicación; para lo cual deberá definir sus propias fronteras de dominio (ver 4.2.4.1 -pág. 10- para ver cómo definir el fichero correspondiente).

3.3 Requisitos específicos para los textos a procesar

YATE sólo puede aplicarse a aquellos textos que hayan sido previamente procesados con las herramientas de análisis lingüístico del IULA. Estos textos pueden ser documentos completos del CT así cómo otros textos procesados en el IULA o usando las mismas herramientas.

En el primer caso (documentos del CT), el programa prevé encontrar los tres ficheros básicos de un documento del CT (*.sgm, *.idx y *.ist) en el directorio c:\tmp y las muestras³ correspondientes en c:\tmp\mostres5. En el segundo caso (otros textos), el programa sólo espera encontrar un fichero con características idénticas a las de una muestra del CT (es decir, sin ficheros básicos).

4 Ejecución

El conjunto de programas y ficheros auxiliares necesarios para la ejecución de *YATE* esta instalados en el servidor Novell del IULA. A pesar de que no es necesario copiarlos al ordenador local se recomienda la creación de un directorio específico desde donde

³ Siempre que en este documento se hace referencia a muestras, se refiere al producto de analizar lingüísticamente un documento (es decir, para los documentos del CT se refiere a los ficheros situados en /usr3/iula/corpus/dominio/mostres5 de los servidores morgana o brangaene).

invocar *YATE*⁴. Así, por ejemplo, el usuario podría crear el directorio `c:\Yate` y por lo tanto para ejecutar el programa debe abrir una ventana de sistema, situarse en el directorio de Yate (`cd \Yate`) y desde allí invocar el programa según se indica más adelante.

A partir de ahora y a efectos de la explicación suponemos que el usuario está trabajando en un entorno Windows2000 y ha creado un directorio en el disco duro local que se denomina “`c:\YATE`” desde donde se ejecutarán todos los mandatos.

Para la obtención de la lista ordenada de CAT es necesario la aplicación de dos programas: `yate.pl` y `yateol.pl`. El primero de ellos obtiene todos los datos básicos (consulta a EWN y aplicación de las diferentes estrategias de extracción individuales sobre el documento considerado) mientras que el segundo se encarga de la combinación de los resultados de cada extractor por medio de diferentes estrategias.

4.1 Procedimiento básico

Para ejecutar el programa es imprescindible abrir una ventana de MS-DOS (icono Símbolo del sistema). Desde esta ventana se ejecutan tanto los programas principales de extracción así como los distintos programas auxiliares. Los programas principales para una ejecución simple de *YATE* son:

```
C:\Yate>perl yate.pl
```

y

```
C:\Yate>perl yateol.pl
```

Podemos obtener una lista completa de todas las opciones disponibles de cada programa invocándolos con la opción `-h`, es decir:

```
C:\Yate>perl yate.pl -h
YATE (Yet Another Term Extractor)
-i docCT          documento SGML (ej. -i m00105.sgm)
-o fileName      nombre base de los ficheros resultado (ej. -o m105)
-lang lang       idioma (es, ca)
-prep prep       preposiciones (+ de 1)
-nogramat       no incluir palabras gramaticales
-noC             trata las conjunciones como frontera
-noArticle      no incluir artículos dintre un CAT
                (impide crear candidatos con el patrón NPAN)
-tag            inclou sec. de tags asociada a cada CT
-ct             limita el cálculo de frecuencia a los CT
-ctx           modo de cálculo del Factor de Contexto
                (global|pond|sem|lex)
-nopr          no calcula precisión y cobertura (aunque encuentre
                los datos)
-npartpr #     número de particiones a usar en cálculo de precisión
                y cobertura
-fastr #       crea los ficheros asociados con FASTR (ej. -fastr
                5000)
-borders file  definición de fronteras de dominio en EWN
-propAdj file  definición de las propiedades a las cuales hacen
                referencia los adjetivos calificativos
-combNA file   definición de las combinaciones admitidas de
```

⁴ Esto es debido a que la ejecución de *YATE* implica la creación de varios ficheros y un subdirectorio en el punto del árbol de directorios desde donde se lo invoca.

Frontera de dominio del Nombre y la Propiedad del
Adjetivo
-stopNSP file fichero con la "stop-list" para la seqüencia Nom+SP
-lemaParticipio file fichero con la informaci³4n para cambiar el lema

de las H

-v verbosidad (ej. -v 1)
-help obtiene este mensaje

Ejemplo:

```
yate.pl -i m00105.sgm -o m105 -nogramat -noC -noarticle -tag -ct -v  
1 -prep de -prep con
```

```
y  
C:\Yate>perl yateol.pl -h  
yateol.pl  
Programa para analizar "offline" los resultados de la ejecución de  
"yate.pl". Versión que actúa a partir del fichero "nom_cat.glob".  
-d nom_documento (nombre base del documento que tiene los patrones)  
-n nickname (nom donat als fitxers generats per excat.pl)  
-lang lang (codi d'idioma: ca|es)  
-p [N NJ NPN] (patró)  
-nopr no calcula los valores de precisión y cobertura  
aunque encuentre los datos  
-npartpr # número de particiones a usar en cálculo de  
precisión y cobertura  
-fastr (opcional: si existe vota fastr)  
-[local|last] (ultimo resultado en directrio de exocat|local)  
-v [0-9] (verbosidad)  
-h (ajuda)
```

Para que *YATE* pueda funcionar es imprescindible la ejecución del procedimiento auxiliar indicado en la sección 5.1. Este procedimiento se debe repetir siempre que modifique alguna de las muestras. En las secciones siguientes se especifica la utilización de cada una de las opciones de cada programa.

4.2 Opciones del programa yate.pl

Este programa se encarga de leer el documento a procesar y sobre él realiza las siguientes tareas:

1. extraer todos los patrones nominales potencialmente terminológicos
2. consultar EWN y calcular el Coeficiente de Especialidad
3. aplicar el método de formantes cultos donde sea posible
4. aplicar el método de uso del contexto
5. crear las páginas web y ficheros de log necesarios para visualizar los resultados de los diferentes análisis realizados
6. creación de la base de datos con los resultados obtenidos

Para realizar estas tareas y modular ciertos aspectos de su funcionamiento el usuario puede indicar una serie de opciones de procesamiento. En esta sección se describe con cierto detalle el funcionamiento de estas opciones.

4.2.1 Documento a procesar

La opción `-d` permite indicar el documento (o fichero) que se desea procesar. Este puede ser un documento del corpus o un documento aislado. En el primer caso, debemos indicar el número de identificación del documento en el CT (ej. `-d m00105.sgm`), en el segundo se debe indicar el nombre del fichero que queremos procesar (ej. `-d fichero.txt`). En el primer caso (documento del CT), la lengua del documento se determina automáticamente mientras que en el segundo (fichero aislado) esta se debe indicar mediante la opción `-lang es` (documento en español) o bien `-lang ca` (documento en catalán).

En la tabla siguiente resumimos y ejemplificamos las diferentes opciones que afectan al documento a procesar:

	Doc. del CT	Doc. aislado
Posición del texto	C:\tmp (ficheros *.sgm, *.idx y *.ist) C:\tmp\mostres5 (muestras)	C:\Yate
Nombre	<code>-d m00105.sgm</code>	<code>-d fichero.txt</code>
Idioma	<code>--</code>	<code>-lang es</code>

4.2.2 Resultados obtenidos

La ejecución de cualquiera de los programas de *YATE* genera una serie de ficheros auxiliares cuya característica común es que su nombre siempre empieza por la cadena que hemos indicado en la opción `-d nombre-salida`. Además, todos estos ficheros se agrupan en un directorio que tiene por nombre la cadena antes indicada.

Podemos acceder a toda la información obtenida a través de una página HTML que tiene por nombre `nombre-salida_main.html`.

Si por ejemplo hemos invocado `yate.pl` de la siguiente manera:

```
C:\Yate>Perl yate.pl -i m00105.sgm -o m105 ....
```

el sistema generará un directorio en `C:\Yate\m105` que contiene una serie de ficheros, uno de los cuales será `m105_main.html`. Esta manera de nombrar los resultados permite realizar la ejecución varias veces, cada una con diferente configuración del módulo de selección de CAT, y después comparar los resultados obtenidos.

4.2.3 Configuración del módulo de selección de CAT

Este módulo se encarga de la selección de CAT, es decir, de aquellas secuencias de texto susceptibles de ser términos. El criterio utilizado para esta selección es estrictamente lingüístico y, cómo en la mayoría de los sistemas de extracción conocidos, se basa en patrones morfológicos. Es decir, segmenta las cadenas en CATs utilizando conocimiento sobre las categorías morfológicas que no pueden formar parte de un término, por ejemplo un nombre puede formar parte de un término mientras que un

determinante o un pronombre no pueden formar parte de un término⁵. Los patrones seleccionados hasta ahora son nombre, nombre-adjetivo y nombre-preposición-nombre.

En consecuencia, este módulo utiliza el conocimiento sobre aquellas categorías morfológicas que nunca podrán formar parte de un término. El criterio es recoger la secuencia de máxima longitud. Los criterios que ha de seguir una secuencia para ser considerada un CAT son los siguientes:

- comenzar con un nombre común
- no incluir pronombres, adverbios ni formas verbales (excepto participios).

Más adelante se muestra cómo el usuario puede configurar el comportamiento de *YATE* en relación a las conjunciones, preposiciones y artículos (en el interior de un CAT).

La configuración de estas opciones tiene consecuencias importantes en los CAT seleccionados por *YATE* y por lo tanto puede modificar los resultados que se obtengan (en particular las cifras de precisión y cobertura). Por ejemplo, un filtro muy restrictivo que no permitiera la selección de CATs que incluyan preposiciones evita la aparición de patrones complejos (NPNJ, NPNPN, etc) pero también la inclusión de CATs con la secuencia NPN; al mismo que maximiza el número de secuencias simples (nombre y nombre-adjetivo). Debe tenerse en cuenta que, de momento *YATE* no trata secuencias complejas (p. ej. NJJ: “circunferencia torácica normal”, NPJJ: “amenaza de parto prematuro” o NPNPN: “presencia de episodio de tos”).

A modo de ejemplo, se presentan tres oraciones y los fragmentos de las mismas que este módulo considera CAT:

- a) *Asma bronquial y reflujo gastroesofágico*
 - *asma bronquial*
 - *reflujo gastroesofágico*

- b) *La curva de respuesta a la histamina en el diagnóstico y tratamiento del asma bronquial*
 - *curva de respuesta*
 - *histamina*
 - *diagnóstico*
 - *tratamiento del asma bronquial*

- c) *La estructuración del primer programa de prevención de enfermedades atópicas (asma bronquial, rinitis alérgica y dermatitis atópica) predispone hacia un cambio de actitud a través de la educación del grupo familiar de riesgo, basado en los resultados de investigaciones previas y adaptado a las normas y costumbres que rigen nuestra sociedad actual.*
 - *estructuración*
 - *prevención de enfermedades atópicas*
 - *asma bronquial*
 - *rinitis alérgica*
 - *dermatitis atópica*
 - *cambio de actitud*
 - *educación del grupo familiar de riesgo*
 - *resultados de investigaciones previas*
 - *normas*
 - *costumbres*
 - *sociedad actual*

⁵ La aproximación seguida para esta selección es similar a la del módulo de delimitación utilizado en el sistema LEXTER de D. Bourigault.

El sistema sólo analizará cómo CAT aquellas secuencias que satisfacen los patrones morfológicos ya indicados. Obsérvese que mientras algunas secuencias son claramente terminológicas (*asma bronquial, reflujo gastroesofágico*) otras pertenecen al lenguaje general (*estructuración, cambio de actitud*). En el ejemplo b) este módulo propone la secuencia *tratamiento del asma bronquial* debido a que es el SN de máxima longitud cuando en realidad el término es *asma bronquial*.

4.2.3.1 Tratamiento de las preposiciones

Por defecto *YATE* considera que todas las preposiciones pueden formar parte de un CAT. Si el usuario desea que el sistema tenga en cuenta sólo una o más preposiciones, debe indicarlo mediante la opción `-prep`. Por ejemplo, si sólo quiere tener en cuenta la preposición “de” deberá añadir la secuencia `-prep de` a la orden básica de procesamiento. Si lo que desea es tener cuenta las preposiciones “de” y “en”, la secuencia a añadir será: `-prep de -prep en`.

Si, en cambio, lo que se quiere es que *YATE* considere la preposición como frontera de término la secuencia a añadir será: `-prep xx`.

Veamos, en un ejemplo real, la lista de CATs seleccionados en función de las preposiciones escogidas para formar parte de un CAT. Consideremos el siguiente oración :

“Con el propósito de comparar el rendimiento de las pruebas de provocación bronquial con ejercicio físico y con histamina en el diagnóstico de asma bronquial se estudiaron 31 niños asmáticos entre 5 y 14 años.”

La tabla siguiente presenta la lista de CATs seleccionados en función del valor asignado a la opción `-prep`⁶.

Opción <code>-prep</code>	CATs seleccionados (sin lematizar)
No especificada	propósito de comparar rendimiento pruebas de provocación bronquial con ejercicio físico histamina diagnóstico de asma bronquial niños asmáticos
<code>-prep de</code>	propósito de comparar rendimiento pruebas de provocación bronquial ejercicio físico histamina diagnóstico de asma bronquial niños asmáticos

⁶ Estos CATs se obtuvieron activando también la opción `-noC` (ver sección 4.2.3.3, pág. 9)

Opción -prep	CATs seleccionados (sin lematizar)
-prep xx	propósito comparar rendimiento pruebas provocación bronquial ejercicio físico histamina diagnóstico asma bronquial niños asmáticos

Es interesante observar cómo aumenta el número de CATs propuestos en función de las preposiciones que se admiten para formar parte de un CAT. En el primer caso, cuando se permite la inclusión de cualquier preposición sólo hay seis CAT, algunos de ellos muy complejos. En el segundo caso, sólo se admite la preposición ‘de’ resultando en siete CATs. Finalmente, en el último caso, no habilita ninguna preposición y obtiene como resultado 10 CATs, todos simples (nombre y nombre –adjetivo). En particular nótese cómo el fragmento “*pruebas de provocación bronquial con ejercicio físico*” se desdobra en dos y tres CATs en función del valor asignado a la opción -prep.

4.2.3.2 Tratamiento de las conjunciones

Por defecto, YATE considera que todas las conjunciones forman parte de un CAT. Si el usuario no desea este comportamiento debe indicarlo mediante la opción -noC.

Veamos el efecto que tiene esta opción en un ejemplo real. Consideremos el fragmento siguiente:

“... con los diagnósticos clínicos de asma bronquial, rinitis alérgica y síndrome bronquial obstructiva recidivante.”

La tabla siguiente presenta la lista de CATs seleccionados en función de si se ha especificado o no la opción -noC⁷.

Opción -noC	CATS seleccionados (sin lematizar)
Especificada	diagnósticos clínicos asma bronquial rinitis alérgica síndrome bronquial obstructiva recidivante
No especificada	diagnósticos clínicos asma bronquial rinitis alérgica y síndrome bronquial obstructiva recidivante

4.2.3.3 Tratamiento de los artículos

No obstante que YATE considera que las unidades terminológicas sólo pueden empezar con un nombre, el usuario puede decidir qué tratamiento desea aplicar a los artículos incluidos en las secuencias nombre-preposición-nombre. Por defecto estas partículas se

⁷ Estos CATs se obtuvieron activando también la opción -prep xx

incluyen en el CAT. Si el usuario no desea este comportamiento debe indicarlo mediante la opción `-noArticle`. Obviamente, esta opción sólo tiene sentido si también se permite que las preposiciones forman parte de un CAT.

Ejemplo, a partir del fragmento de texto siguiente:

... prevención de las enfermedades atópicas ...

el comportamiento del módulo extractor de candidatos podría proponer los CAT siguientes:

Opción <code>-noArticle</code>	CATs seleccionados (sin lematizar)
Especificada	prevención enfermedades atópicas
No especificada	prevención de las enfermedades atópicas

4.2.4 Adaptación de *YATE* a diferentes dominios

YATE es un sistema de extracción de terminología desarrollado para el dominio médico. A pesar de esto, es posible adaptar este sistema a cualquier otro dominio a condición de que se pueda determinar el contenido de algunos ficheros de configuración. Más concretamente se han de definir las siguientes informaciones relevantes al dominio que se quiere definir:

- las fronteras de dominio,
- las propiedades relevantes,
- las combinaciones pertinentes nombre + adjetivo calificativo y
- la lista de exclusión para las secuencias NPN.

En las cuatro subsecciones siguientes se especifican el formato y contenido de estos cuatro ficheros⁸. También se incluirá información sobre cómo se le indica a *YATE* que utilice los nuevos ficheros en lugar de los que accede por defecto. Para poder definir el contenido de la mayoría de ficheros indicados es imprescindible tener acceso a EWN⁹. En [Vivaldi, 2001] se justifica y ejemplifica el uso de esta información.

4.2.4.1 Definición del dominio

YATE considera que un dominio está definido mediante la enumeración de una serie de synsets de EWN junto con cierta información añadida para cada uno de ellos. *YATE* supone que estos synsets (y las variantes asociadas) junto con todos los que son directa o indirectamente sus hipónimos son relevantes al dominio que se está definiendo.

Por defecto, *YATE* considera fronteras de dominio todos los synsets incluidos en el fichero `DomainMedic` y que pretenden definir al dominio de la Medicina. Uno de las fronteras de dominio incluidas en este fichero está definida por el synset `08587853n` que corresponde a la palabra (variante en el léxico propio de EWN) “enfermedad”. La línea correspondiente tiene el aspecto siguiente:

```
08592183n Disease 4 00015437n
```

Al escoger este synset como frontera de dominio *YATE* está considerando a este synset así como a todos su hipónimos (1337) como pertenecientes al dominio médico.

⁸ Recomendamos el uso de un simple editor de textos (p.ej. El bloc de notas) para editar estos ficheros. Es posible, pero no es conveniente, utilizar un procesador de texto (ej. Microsoft Word).

⁹ Una posible vía de acceso es a través de <http://nipadio.lsi.upc.es/cgi-bin/public/wei2.consult.perl>

Si queremos definir nuestras propias fronteras de dominio debemos crear un nuevo fichero similar a `DomainMedic` e indicarle a *YATE* dicha situación mediante la opción `-borders`. Por ejemplo, si hemos definido las fronteras del dominio del Genoma en el fichero `genomicsBorder` debemos invocar `yate.pl` de la siguiente manera:

```
C:\Yate>Perl yate.pl -i ... -o ... -borders DomainMedic ...
```

El usuario deberá crear el fichero `genomicsBorder` en su ordenador local y en el directorio destinado a *YATE* (por ejemplo en `c:\Yate`). A continuación describimos el formato que debemos dar a dicho fichero.

El fichero de definición de dominio estará compuesto por una o más líneas cada una de las cuales corresponderá a una frontera de dominio. El formato de estas línea debe ser el siguiente:

```
[synset]n\t[tag]\t[distancia al top]\t[synset del top]n
```

donde:

<code>synset</code>	número de identificación del synset que consideramos frontera
<code>tag</code>	etiqueta arbitraria y única que asignamos al synset
<code>distancia al top</code>	distancia, en número de synsets, para llegar al nodo superior (o top) de la jerarquía donde se encuentra el synset.
<code>synset del top</code>	número de identificación del synset que hace de top
<code>\t</code>	tabulador

En este fichero se consideran líneas de comentario todas aquellas que empiezan con la secuencia `##`. A continuación se muestra, a modo de ejemplo, un fragmento del fichero `DomainMedic`.

```
## Información sobre los "ilis" de EWN que pertenecen ## al
## dominio médico
## Separador de campo: 1 (o más) tabuladores (\t)
## 0: ili "médico"
## 1: tag semántico
## 2: distancia al top
## 3: ili del top

08592183n Disease 5 00015437n
08586618n Malfunction 2 00015437n
08767168n Predisposition 3 00015437n
08586350n HealthProblem 3 00015437n
08577911n PhysiologicalState 1 00015437n
...
```

4.2.4.2 Definición de las propiedades relevantes para los adjetivos calificativos pertinentes en un dominio

EWN considera que todo adjetivo calificativo da valor a una propiedad. *YATE* utiliza esta peculiaridad para seleccionar aquellas secuencias nombre + adjetivo calificativo que son pertinentes en medicina. Para ello es necesario crear un archivo de texto donde estén indicados todas las propiedades relevantes en el dominio que estamos analizando.

En el dominio médico, por ejemplo, el synset `03562490n` que tiene cómo variantes las cadenas “amplitud”, “anchura” y “ancho” corresponde a una propiedad que indica la

extensión de un objeto¹⁰. Al escoger este synset cómo propiedad relevante estamos haciendo que *YATE* pueda reconocer términos cómo “tórax estrecho”. La línea correspondiente tiene el aspecto siguiente:

```
03562490n Anchura 6 00012670n
```

Por defecto *YATE* considera que son propiedades relevantes las que están incluidas en el fichero `DomainMedicAdj` que corresponden al dominio de la Medicina. Si queremos definir nuestras propias propiedades y que estas sean utilizadas por *YATE* debemos indicarlo por medio de la opción `-propAdj`. Por ejemplo, si hemos definido las propiedades del dominio del Genoma en el fichero `genomicsProperties` debemos invocar `yate.pl` de la siguiente manera:

```
C:\Yate>Perl yate.pl -i ...-o ...-propAdj genomicsProperties ...
```

Este fichero estará compuesto cómo mínimo por una serie de líneas cada una de las cuales corresponderá a una propiedad relevante dentro del dominio de interés. El formato de estas líneas debe ser el siguiente:

```
[synset]n\t[propiedad]\t[distancia al top]\t[synset del top]n
```

donde:

synset	número de identificación del synset
propiedad	etiqueta arbitraria y única que asignamos al synset
distancia al top	distancia en número de synsets para llegar al nodo superior (o top) de la jerarquía donde se encuentra el synset
synset del top	número de identificación del synset que hace de top
\t	tabulador

En este fichero se consideran líneas de comentario todas aquellas que empiezan con la secuencia ‘##’. A continuación se indica, a modo de ejemplo un fragmento del fichero `DomainMedicAdj`.

```
## Información sobre los "ilis" de EWN que corresponden
## a PROPIEDADES relevantes al dominio médico
## Separador de campo: 1 o más tabuladores (\t)
## 0: ili que corresponde a una "propiedad médica"
## 1: tag de la propiedad
## 2: distancia al top
## 3: ili del top
```

```
50003947n Cronicidad 3 00012670n
03562490n Anchura 6 00012670n
03534490n PosicionVertical 5 00012670n
...
```

4.2.4.3 Especificación de las combinaciones pertinentes de nombre + adjetivo calificativo

YATE considera que una secuencia nombre + adjetivo calificativo es relevante en un dominio cuando la combinación de frontera médica del nombre y la propiedad a la cual el adjetivo calificativo da valor ha sido declarada cómo relevante. Para ello es necesario

¹⁰ La definición exacta es la siguiente: “*the extent of something from side to side*”.

crear un fichero de texto donde estén indicadas las combinaciones válidas para el dominio considerado.

Por ejemplo, consideremos el término “tórax estrecho” que consideramos pertinente en el ámbito de la Medicina. El primer elemento de este término es el nombre “tórax” que tiene por frontera de dominio el synset con la etiqueta “bodyPart”. El segundo elemento es el adjetivo “estrecho” que es un valor de la propiedad “anchura”. Si de alguna manera indicamos que la combinación “bodyPart” y “anchura” da lugar a una unidad terminológica relevante en el dominio médico, *YATE* será capaz de reconocer la secuencia “tórax estrecho” como un término. La línea del fichero de texto correspondiente tiene la forma siguiente:

```
03610098n BodyPart 03562490n Anchura
```

Evidentemente tanto la frontera de dominio (BodyPart) como la propiedad (Anchura) deben declararse en los ficheros correspondientes (ver secciones 4.2.4.1 y 4.2.4.2).

Por defecto, *YATE* considera como combinaciones relevantes las que están incluidas en el fichero DomainCombineNA que corresponden al dominio de la Medicina. Si queremos definir nuestras propias propiedades debemos indicarlo por medio de la opción `-combNA`. Por ejemplo, si hemos definido las combinaciones frontera-propiedad del dominio Genoma en el fichero `genomicsCombinesNA` debemos invocar `yate.pl` de la siguiente manera:

```
C:\Yate>Perl yate.pl -i ...-o ...-combNA genomicsCombinesNA ...
```

Este fichero estará compuesto como mínimo por una serie de líneas cada una de las cuales corresponderá a una propiedad relevante dentro del dominio de interés. El formato de estas líneas debe ser el siguiente:

```
[synset1]n\t[frontera de dominio]\t[synset2]n\t[propiedad]n
```

donde:

synset1	número de identificación del synset que actúa como frontera
frontera de dominio	etiqueta de frontera que asignamos al synset1
synset2	número de identificación del synset que identifica la propiedad
propiedad	etiqueta arbitraria con la que identificamos al synset2
\t	tabulador

Se consideran líneas de comentario todas aquellas líneas de este fichero que empiezan con la secuencia ‘##’. A continuación se indica, a modo de ejemplo un fragmento del fichero `DomainCombineNA`.

```
## Información sobre las características semánticas que
## hacen que un CAT con patrón nombre-adjetivo sea válido.
## Separador de campo: uno o más tabuladores (\t)
##
## 1) synset+pos de la frontera médica
## 2) etiqueta (semántica) de esa frontera médica
## 3) synset+pos de la propiedad (que es valorada por el
##    adjetivo calificativo)
## 4) etiqueta (semántica) de esa propiedad

## Acepta casos como "asma aguda", "tos crónica", ...
```

```
08592183n Disease 50003947n Cronicidad
08671032n Symptom 50003947n Cronicidad

## Acepta casos como "torax estrecho", ...
03610098n BodyPart 03562490n Anchura
...
```

4.2.4.4 Lista de exclusión para las secuencias NPN

El tratamiento de las secuencias nombre-preposición-nombre es extremadamente complejo por lo que el tratamiento que hace *YATE* de esta secuencia es muy superficial. Uno de los aspectos que se tiene en cuenta para desechar un CAT es cuando el primer nombre es un paratérmino, es decir sirve para dar soporte a un término. Así por ejemplo, la secuencia “caso de asma” no será aceptada como válida ya que no aporta información específica del dominio. Algunos ejemplos similares de paratérminos son los siguientes:

paratérmino	CATS descartados
interior	interior del estómago, interior del pulmón,...
presencia	presencia de síntomas, presencia de mixomas,...
estado	estado de inflamación, estado de salud,...
aparición	aparición de tos, aparición de sibilancias,...

Por defecto, *YATE* considera como lista exclusión los nombres que están incluidos en el fichero `stoplist_NSP` que han sido definidos para el dominio de la Medicina. Si queremos definir nuestra propia lista de exclusión debemos indicarlo por medio de la opción `-stopNSP`. Por ejemplo, si hemos definido una lista de exclusión que consideramos específica del dominio Genoma en el fichero `genomicsStopNSP` debemos invocar `yate.pl` de la siguiente manera:

```
C:\Yate>Perl yate.pl -i ...-o ...-stopNSP genomicsStopNSP ...
```

Este fichero tiene un formato muy simple ya que es una enumeración de todos los lemas que se desea incluir. Al igual que en el resto de ficheros de configuración, el sistema considera líneas de comentario todas aquellas líneas de este fichero que empiezan con la secuencia ‘##’.

4.3 Opciones del programa `yateol.pl`

Este programa se encarga fundamentalmente de ordenar, según diferentes criterios, los resultados obtenidos mediante la combinación de estrategias por votación y también de presentar los resultados para algunos métodos de extracción de términos. La lista de tareas que realiza este programa se puede resumir como sigue:

1. leer los resultados obtenidos por la aplicación del programa `yate.pl`
2. combinar (por votación) los resultados obtenidos por los métodos individuales.
3. obtener los valores de precisión y cobertura (sólo si dispone de la lista de términos validados para los patrones nombre, nombre-adjetivo y nombre-preposición-nombre)
4. crear las páginas web y ficheros de log necesarios para completar la visualización de los resultados de los diferentes análisis realizados

Para realizar estas tareas y modular ciertos aspectos de su funcionamiento el usuario dispone de una serie de opciones de procesamiento. En esta sección se describe con cierto detalle el funcionamiento de estas opciones.

4.3.1 Documento a procesar

La opción `-d` de este programa es muy similar a la del programa `yate.pl` es decir permite indicar el documento que se desea procesar. Este puede ser un documento del corpus o un documento aislado. En el primer caso debemos indicar el número 'sgm' del documento (ej. `-d m00105.sgm`) mientras que en el segundo lo que se debe indicar es el nombre del fichero que queremos procesar (ej. `-d fichero.txt`). En el primer caso la lengua del documento se determina automáticamente mientras que en el segundo esta se debe indicar mediante la opción `-lang es` (documento en español) o bien `-lang ca` (documento en catalán).

La opción `-n` indica de dónde deben tomarse los resultados de la etapa de exploración previa efectuada por el programa `yate.pl`. Por lo tanto, debe especificarse de la misma manera que la opción `-o` utilizada para la ejecución de `yate.pl` que interesa utilizar.

En la tabla siguiente resumimos las diferentes opciones que afectan al documento a procesar:

	Doc. del CT	Doc. Aislado
Opción <code>-o</code> de <code>yate.pl</code>	<code>-o m105</code>	<code>-o fichero</code>
Posición del texto	C:\tmp (ficheros *.sgm, *.idx y *.ist) C:\tmp\mostres5 (muestras)	C:\Yate
Nombre	<code>-d m00105.sgm</code>	<code>-d fichero.txt</code>
Idioma	--	<code>-lang es</code>
Origen de los datos	<code>-n m105</code>	<code>-n fichero</code>

4.3.2 Patrón a procesar

Esta opción permite especificar qué patrón se desea procesar. Es decir, permite actualizar los resultados para las secuencias nombre (`-p N`), nombre-adjetivo (`-p NJ`) o nombre-preposición-nombre (`-p NPN`). Si el usuario no indica ningún patrón en concreto, el programa procesará los tres antes indicados (N, NJ y NPN).

4.3.3 Cálculo de precisión y cobertura

Es posible evaluar el comportamiento de *YATE* en relación con la ordenación de los CATs extraídos para un cierto documento. Esta evaluación se realizará utilizando el cálculo estándar de precisión y cobertura. Para poder efectuar dicho cálculo es imprescindible disponer de la lista de términos validados para dicho documento. Para ello será necesario que un especialista en el dominio al que pertenece el documento

procesado evalúe cuidadosamente todo el documento y extraiga la lista de términos válidos¹¹.

Si el programa `yateol.pl` detecta la presencia de la lista de términos válidos para el documento automáticamente realizará los cálculos de precisión y cobertura y generará las gráficas correspondientes. En principio, este cálculo se realiza dividiendo el conjunto de términos válidos en diez fragmentos iguales. Hay dos opciones de procesamiento que permiten alterar este modo de funcionamiento:

- `nopr` Inhibe el cálculo de precisión y cobertura aunque encuentre la información necesaria para ello.
- `partpr n` Establece `n` cómo el número de particiones a utilizar en el cálculo de precisión y cobertura (por defecto es 10).

La lista de términos válidos se deberá procesar como se indica en la sección 5.2.

4.4 Visualización de los resultados

YATE produce un conjunto de ficheros, algunos de los cuales informan sobre los resultados alcanzados mientras que otros recogen datos intermedios que sólo son útiles para comprobar el funcionamiento interno. Todos los datos relativos a los resultados pueden verse en forma de páginas web mediante ficheros que se han generado en el subdirectorio indicado por la opción `-o` (de `yate.pl`).

Podemos acceder a todas las páginas de resultado abriendo el fichero `(salida)_main.htm` con un visualizador de páginas web. `salida` es el valor dado a la opción `-o`. Si por ejemplo, hemos invocado al extractor con:

Extracció de Candidats a Terme a m105 (30/04/2003)

Dades del document
Datos bibliográficos del documento analizado

Resum de la cerca a EWN
Resultados de la búsqueda de nombres y adjetivos en EWN.

Condicions de la cerca
Especifica las principales opciones dadas al programa yate.pl (fronteras de dominio, propiedades, configuración del módulo de selección de CATs) así como el número de términos validados para cada patrón.

Llista de lemes
Enlace a una página web donde se muestra la lista de los lemas presentes en el documento explorado

Anàlisi dels Candidats a Terme
Enlace a una página web que muestra para cada patrón todos los CAT y para cada uno de ellos toda la información asociada (frecuencia, Coeficiente de Especialidad, reconocimiento por Formantes Cultos, información estadística, contextos, etc.

¹¹ En caso de realizar una consulta a un especialista se recomienda explicar con sumo cuidado y detenimiento la definición de término, el objetivo del experimento así cómo la necesidad de indicar “todos” los términos.

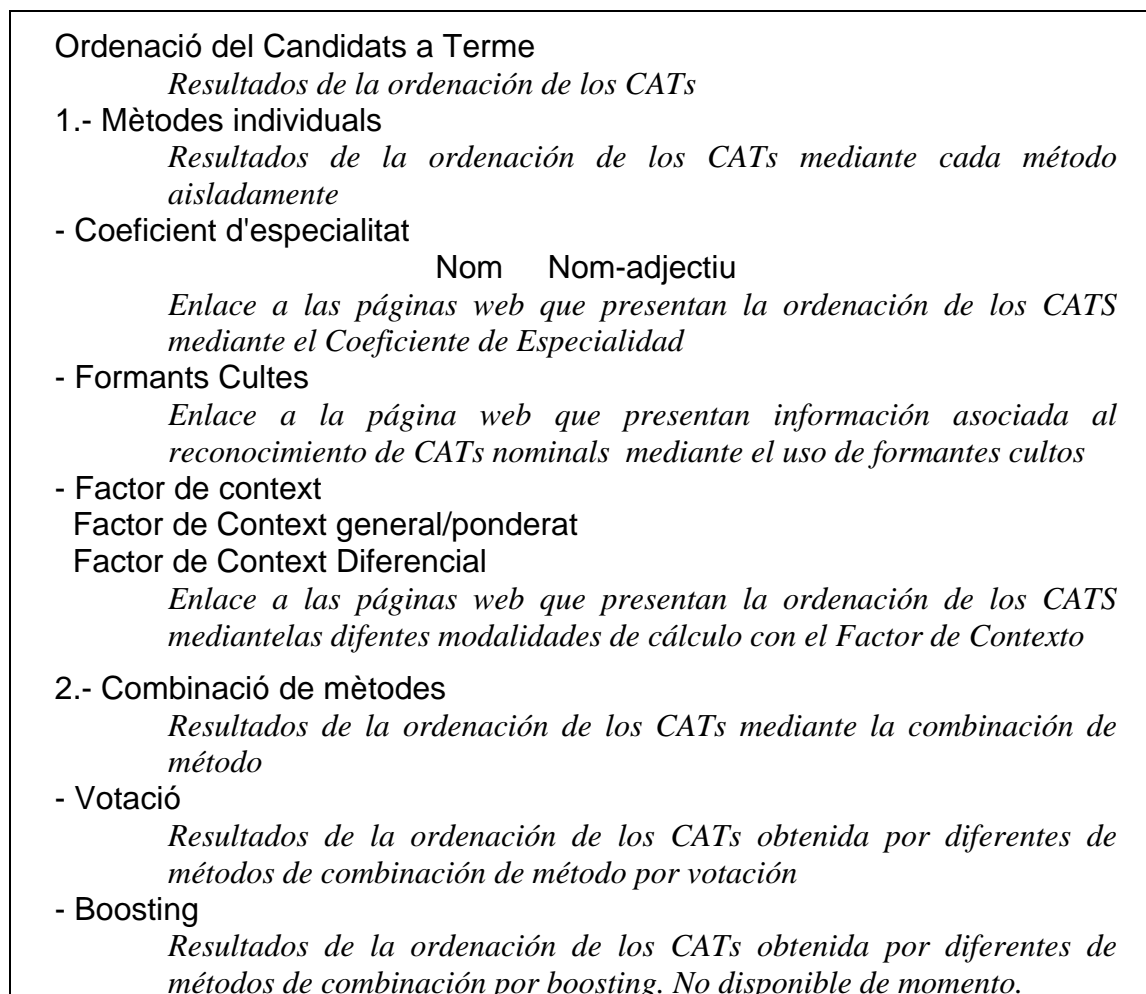


Figura 2. Estructura de la página principal de visualización de los resultados obtenidos por la aplicación de YATE

```
C:\Yate>Perl yate.pl -i m00105.sgm -o m105 ....
```

Debemos abrir el fichero localizado en C:\Yate\m105\m105_main.htm

A partir de esta página se puede acceder a toda la información recogida por YATE sobre el documento analizado. La figura 1 muestra la estructura de esta página de datos.

4.5 Limitaciones

Aunque YATE es un sistema de extracción de CAT nominales, en su estado de desarrollo actual, no puede tratar todos los fenómenos asociados a la aparición de estas unidades en textos escritos.

- No trata la coordinación a la derecha o a la izquierda. Un CAT como “niños y adultos asmáticos” debería dar lugar a dos candidatos separados: “niños asmáticos” y “adultos asmáticos”. Otros ejemplos similares son: “células inmunes y accesorias”; “frecuencia cardiaca y respiratoria”; “formas leves, moderadas y graves”, “episodios ocasionales de disnea y sibilancias”, etc.

- Sólo trata candidatos que respondan a los patrones nombre, nombre-adjetivo o nombre-preposición-nombre. El sistema ignora el resto de patrones sin intentar segmentar secuencias complejas como por ejemplo: “*causa frecuente de síntoma de alergia respiratorio*”.
- No puede detectar términos que no sean secuenciales. Por ejemplo, no detecta el término: “*leucocitosis polimorfonuclear*” en la frase: “*En las pruebas complementarias existe leucocitosis con predominio polimorfonuclear, más acentuada durante los episodios de fiebre.*”

5 Procedimientos auxiliares

5.1 Obtención del fichero índice

El fichero índice obtiene los datos que permitirán a `yate.pl` obtener la información de contexto de todos los CAT. La orden básica para obtener este fichero es la siguiente:

```
perl k:\utils\indexsgmlnew.pl
  -i documento
  -o fichero_de_salida.5dx
```

Si queremos obtener el fichero índice para el documento `m00105` el mandato a ejecutar es el siguiente:

```
C:\Yate>perl k:\utils\indexsgmlnew.pl -i m00105.sgm -o m00105.5dx
```

Es muy importante tener en cuenta que siempre que de alguna manera se modifique alguna muestra debe volver ejecutar este procedimiento antes de ejecutar `yate.pl`.

5.2 Datos necesarios para el cálculo de precisión y cobertura

YATE puede calcular la información de precisión y cobertura de una extracción dada si previamente se ha creado una base de datos con los términos presentes en el documento que se quiere procesar. Obviamente `yate.pl` supone que todos y cada uno de los términos presentes en el documento están incluidos en esta base de datos. Este cálculo se limita aquellos términos que correspondan a los patrones nombre, nombre-adjetivo y nombre-preposición-nombre.

Para obtener la base de datos antes mencionada es necesario crear un directorio denominado `DatosPR` en el directorio base de procesamiento (`c:\Yate`) e incluir en él tres ficheros de texto con los términos de cada uno de los patrones. El nombre de cada uno de estos ficheros debe tener el formato siguiente:

`documento_patrón.txt`

Cada uno de estos ficheros debe tener un término por línea y una primera línea que indique el patrón y el documento que se trate. Esta primera línea debe seguir el formato siguiente:

`Document:\tdocumento\tPatrón: *patrón*`

Ejemplo: los términos del patrón nombre del documento `m00105.sgm` están incluidos en un fichero denominado `m00105_N.txt` incluido en el directorio `c:\Yate\DatosPR`. Reproducimos a continuación un fragmento de dicho fichero:

```
Document: m00105.sgm      Patrón: *N*
ablactación
```

aborto
absorción
acetilcolina
acidosis
acné
acupuntura
adenoamigdalitis
...

Una vez que se han creado los ficheros con los términos presentes en el documento debe ejecutarse el programa `creaPRdb.pl`. Este programa sólo admite como parámetro la opción `-d` que permite indicar el documento del cual queremos crear la base de términos. Ejemplo: para el documento `m00105` debemos crear los ficheros (en `c:\Yate\DatosPr`) `m00105_N.txt`, `m00105_NJ.txt` y `m00105_NPN.txt`. y después ejecutar:

```
C:\Yate>k:\Utils\creaPRdb.pl -d m00105
```

que creará el fichero `m00105_terms.dat` (en `C:\Yate`).

6 Referencias

Vivaldi Jorge (2001) “Extracción de candidatos a término mediante combinación de estrategias heterogéneas”. Tesis Doctoral. Universidad Politécnica de Catalunya.