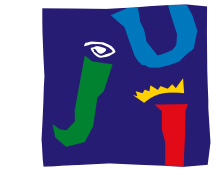


Jornades de Foment de la Investigació

ETIQUETADO DE TEXTOS CONCONNEXOR

Autor

Samia HOMSANI.

RESUMEN

Este artículo trata sobre el etiquetado morfosintáctico de textos electrónicos mediante la utilización de un programa de etiquetado. Para llevar a cabo el proceso se ha utilizado el programa *Connexor Machinese Syntax*; los textos etiquetados pertenecen a los corpus TXTCeram [1] y TXTInfo-ES.

El etiquetado de los corpus TXTCeram y TXTInfo-ES se enmarca dentro del proyecto ONTODIC , cuyo objetivo general es la creación de un diccionario terminológico onomasiológico. Además, otros de los objetivos del proyecto versarán acerca de la extracción terminológica y la extracción de información conceptual. El proyecto ONTODIC[2] ha sido concebido por el grupo de investigación TecnoleTTra (tecnologías del lenguaje, la terminología y la traducción) de la Universitat Jaume I, y está dirigido por Amparo Alcina.

El marcado o etiquetado de textos electrónicos consiste en codificar documentos con la información morfosintáctica referente a cada una de las unidades por las que están formados, de forma que éste pueda ser posteriormente procesado por programas informáticos.

La herramienta *Connexor Machinese Syntax* nos permite la realización automática del etiquetado, ya que procesa los archivos en formato .txt, y nos permite obtener los resultados del análisis morfosintáctico tanto en formato de texto etiquetado como en un documento etiquetado en xml.

[1] TXTCeram: Extracción semiautomática y análisis conceptual formal de términos de la cerámica a partir de un corpus electrónico. Su eficacia y utilidad en la mediación lingüística es un proyecto financiado por la Generalitat Valenciana (GV05/260).

[2] ONTODIC: Metodología y tecnologías para la elaboración de diccionarios onomasiológicos basados en ontologías. Recursos terminológicos para la e-traducción en un proyecto financiado por el Ministerio de Educación y Ciencia de España (TSI2006-01911).

1. INTRODUCCIÓN

Durante el presente trabajo se explicará cuál ha sido el proceso de etiquetado de los corpus TXTCeram y TXTInfo-ES mediante la puesta en práctica de una herramienta informática de etiquetado morfosintáctico. Con el fin de aclarar algunos conceptos generales, se explicará qué es un corpus y cuáles son sus principales características y se introducirá brevemente en qué consiste el etiquetado automático de textos.

La finalidad primera del trabajo consistía en obtener un corpus lingüístico en español etiquetado morfosintácticamente para poder aplicarlo en el proyecto ONTODIC, que tiene como meta la elaboración de una ontología de la cerámica que combine características de las ontologías lingüísticas y ontologías de dominio. No obstante, el corpus resultante será utilizado para efectuar una extracción terminológica, para la extracción de información referente por ejemplo a las relaciones conceptuales, y también para establecer una comparación entre el etiquetado de los corpus que nos permitirá alcanzar conclusiones acerca de los resultados derivados del proceso según el grado de especificidad de los textos que conforman ambos corpus.

2. CORPUS LINGÜÍSTICOS, ELECTRÓNICOS Y ETIQUETADOS

Un corpus es un conjunto de textos que se han seleccionado previamente según ciertos criterios, y que pueden servir a una investigación lingüística y traductológica. Los corpus pueden clasificarse teniendo en cuenta diferentes características: según la lengua, es decir, si son monolingües, bilingües o multilingües; según si es cerrado o abierto (si se le añaden textos progresivamente o si al alcanzar un número determinado de palabras ya no se le añaden más, por ejemplo); según el tipo de textos que incluye; según si es crudo (textos originales sin información añadida) o etiquetado (si se le añade información lingüística referente, por ejemplo, a la estructura, la morfología, la sintaxis, etc.). Asimismo, el tipo de corpus también depende del modo en que podemos realizar la consulta lingüística. En algunos casos, para consultar un corpus tendremos que recurrir a la forma impresa y en otros casos, podrá ser una consulta en línea.

Algunos de los corpus que se pueden consultar en línea son el corpus técnico etiquetado del Instituto Universitario de Lingüística Aplicada, IULA, (http://bwananet.iula.upf.edu/bwananet1a.ca.htm) de la Universidad Pompeu Fabra y el CREA, el corpus de referencia del español actual de la Real Academia Española (http://corpus.rae.es/creanet.html).

Como se ha mencionado antes, un corpus etiquetado está formado por textos codificados con información lingüística en distintos niveles (documental, textual, morfológica, sintáctica, etc.) mediante etiquetas (Alcina, 2005). Las etiquetas, que suelen aparecen precedidas por un delimitador (ej. &, @, <>, etc.), pueden identificar información morfológica como, por ejemplo, el género masculino o femenino de un sustantivo o un adjetivo. Por lo tanto, el marcado de textos electrónicos consiste en codificar información referente al contenido del documento de forma que este pueda ser posteriormente procesado por programas informáticos.

3. EL ETIQUETADO

En el proceso de etiquetado automático de textos, la herramienta informática procesa los diferentes textos pertenecientes al corpus para obtener el análisis del documento. Durante este proceso, el programa marca mediante etiquetas las características de las palabras según los criterios de clasificación de la información que deseemos.

Normalmente, en cualquier ejemplo de texto etiquetado encontraremos una clasificación por columnas; así, cada columna nos dará una información diferente: la palabra tal y como aparece en el texto, la base léxica de la palabra, la relación sintáctica entre la palabra y el elemento de la frase al que modifica y la etiqueta precedida, como hemos apuntado antes, por un delimitador. Un ejemplo de texto etiquetado morfológica y sintácticamente sería:

#	Text	Baseform	Syntactic relation	Syntax and morphology
1	Ana	ana	subj:>2	&NH N FEM SG
				&NH <Proper> N FEM SG
2	Compró	comprar	main:>0	&+FM V IND PRET SG3
3	cinco	cinco	qn:>4	&QN> NUM CARD
4	manzanas	manzana		&NH N FEM PL
5	rojas	Rojo	ads:>4	&<A A FEM PL
6	para	Para	pm:>7	&PM> PREP
7	Juan	Juan		&NH <Proper> N MSC SG

Ejemplo de texto etiquetado con la herramienta STILUS

Entre las diferentes herramientas informáticas que nos permiten etiquetar un texto de forma automática se encuentran tanto programas con licencia como etiquetadores *on line*. Son etiquetadores *on line* el etiquetador morfosintáctico de STILUS, perteneciente al grupo Daedalus, una empresa dedicada al sector de las Tecnologías de la Información y de las Comunicaciones; y el etiquetador del Grup d'Investigació en Lingüística Computacional de la Universitat de Barcelona. El primer etiquetador sólo permite etiquetar en español y nos permite desambiguar morfológicamente. El segundo, no nos devolvía la oración etiquetada.

Dentro de los etiquetadores con licencia están el TreeTagger, desarrollado dentro del proyecto TC de la Universidad de Stuttgart, o *Connexor*, ambos son anotadores morfosintácticos multiligües.

3.1. La herramienta de etiquetado Connexor

Después de haber comprobado las propiedades de los programas on line y los programas con licencia se optó por utilizar uno de estos últimos por una cuestión de calidad. El grupo Tecolettra decidió utilizar una herramienta que etiquetara los textos de manera automática y marcara tanto la información morfológica como

las relaciones sintácticas de los textos procesados. Treetagger y *Connexor* nos ofrecían estas propiedades, sin embargo, tras tener varias entrevistas con grupos de investigación de otras universidades que ya habían trabajado con Treetagger, la herramienta seleccionada fue *Connexor Machinese Syntax*.

El programa de pago *Connexor*, y más concretamente su herramienta *Connexor Machinese Syntax* (http://www.connexor.com), ofrece la obtención de la información representada en las frases del documento mediante un análisis morfológico así como de las relaciones sintácticas existentes entre los tokens.

Además, permite el análisis de textos en diferentes idiomas, como pueden ser el inglés, el alemán, el francés, el español o el sueco. Teniendo en cuenta que el número de etiquetadores para otras lenguas que no sean el inglés es muy reducido hay que considerar que *Connexor Machinese Syntax* intenta proporcionar unos resultados lo más uniformes posible sin dejar de prestar atención a las especificidades de cada lengua.

Connexor Machinese Syntax marca los textos basándose principalmente en cinco campos:

- la posición de las palabras en la oración
- la palabra
- el tipo de raíz
- la relación sintáctica
- las etiquetas sintácticas y las morfológicas

Además, dentro de cada categoría gramatical (sustantivos, verbos, adjetivos, pronombres, determinantes, numerales, adverbios, preposiciones, conjunciones, interjecciones, abreviaturas…), *Connexor Machinese Syntax* etiquetará las palabras según el género, número, modo, persona, tiempo, etc.

Si tenemos en cuenta lo costoso que puede resultar recabar información mediante el análisis de corpus extensos como los de TXTCeram y TXTInfo-ES, es evidente que el programa *Connexor Machinese Syntax Client* es la herramienta idónea para realizar esta tarea ya que, entre otras funciones, nos permite analizar textos transformándolos en documentos cuyos datos y estructura son analizados lingüísticamente; además, y considerando que el grupo TecnoleTTra trabaja con textos en español y en inglés, es importante que *Connexor Machinese Syntax Client* nos permita trabajar de forma multilingüe.

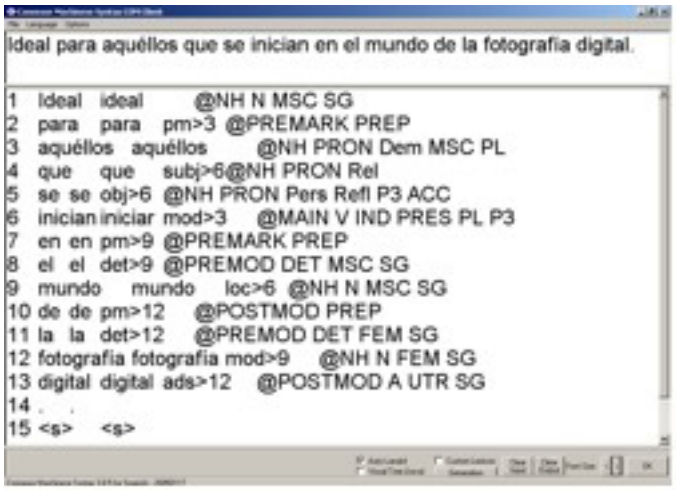
El etiquetado de las frases proporciona información suficiente como para llevar a cabo diferentes tareas de procesamiento de la información, como por ejemplo, la búsqueda de las palabras clave y la extracción terminológica.

Al mismo tiempo, el hecho de que el programa marque las relaciones sintácticas existentes entre los diferentes elementos de la frase es de gran utilidad en el tema que nos ocupa ya que se puede realizar una extrac-

ción terminológica uniendo palabras o cadenas de palabras según su categoría morfológica o sus relaciones sintácticas.

3.2. Las etiquetas de Connexor Machinese Syntax Client

A continuación, se explicará con detalle cómo *Connexor Machinese Syntax* marca los textos:



Anteriormente se ha mencionado que Connexor Machinese Syntax se basa en cinco campos. De este modo, a los lados de cada palabra aparecen sus etiquetas correspondientes indicando el análisis gramatical realizado:

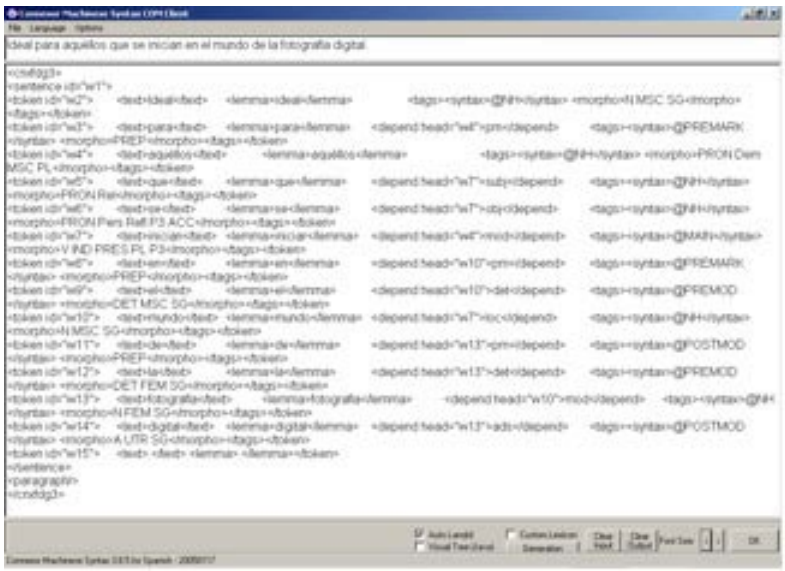
El primer campo, la posición de las palabras en la oración, está marcado por los números alineados en la columna de la izquierda, así, en este ejemplo, la oración está formada por 15 elementos. En la segunda columna, podemos ver las palabras que forman la oración (ej. "ideal", "en", "fotografía", "digital"…). La tercera columna, nos muestra la raíz de la palabra (ej. "inician" cuya raíz es "iniciar"). La cuarta columna designa la relación sintáctica de cada elemento con respecto al resto de elementos de la frase, por ejemplo la palabra "la" es un determinante (DET) femenino (FEM) singular (SG) prenominal (PREMOD) que modifica a la palabra perteneciente al número 12: "fotografía" (det>12). Por último, en la quinta columna, precedida por el delimitador, una arroba, incluye todas las marcas morfológicas, así "inician" cuya raíz es "iniciar" y que acompaña al elemento "aquéllos" es el verbo (V) principal (MAIN) de la oración; está en presente (PRES) de indicativo (IND), tercera persona (3P) del plural (PL).

En la siguiente tabla se puede observar un resumen del análisis que Connexor ha hecho de la oración:

Texto etiquetado	Información proporcionada por las etiquetas
1 ideal Ideal @NH N MSC SG	Sustantivo masculino singular
2 para para pm>3 @PREMARK PREP	Preposición
3 aquéllos aquéllos @NH PRON Dem MSC PL	Pronombre demostrativo masculino plural
4 que que subj>6 @NH PRON Rel	Pronombre relativo
5 se se obj>6 @NH PRON Pers Refl P3 ACC	Pronombre personal reflexivo tercera persona del plural
6 inician iniciar mod>3 @MAIN V IND PRES PL P3	Verbo principal, presente de indicativo tercera persona del plural
7 en en pm>9 @PREMARK PREP	Preposición
8 el el det>9 @PREMOD DET MSC SG	Determinante masculino singular
9 mundo mundo loc>6 @NH N MSC SG	Sustantivo masculino singular
10 de de pm>12 @POSTMOD PREP	Preposición
11 la la det>12 @PREMOD DET FEM SG	Determinante femenino singular
12 fotografía fotografía mod>9 @NH N FEM SG	Sustantivo femenino singular
13 digital digital ads>12 @POSTMOD A UTR SG	Adjetivo singular

Además, cabe destacar que con *Connexor Machinese Syntax* podemos etiquetar los textos en formato .xml (eXtensible Mark up Language). El lenguaje XML, una adaptación del SGML, está considerado como extensible ya que permite al usuario definir sus propias etiquetas. Es un metalenguaje estándar que define la gramática de lenguajes específicos y que fue desarrollado por el World Wide Web Consortium (un consorcio internacional que crea lenguajes para la World Wide Web). El XML se plantea como un estándar que permite el intercambio de información estructura en diferentes plataformas y además de poder ser utilizado en aplicaciones informáticas, puede ser aplicado en bases de datos, editores de texto, hojas de cálculo, etc. Por ello, el carácter extensible de XML nos permite compartir la información estructurada entre diferentes sistemas de una manera sencilla y segura. Hemos visto, que en algunos casos los delimitadores de las etiquetas eran símbolos como el *ampersand* (&) o la arroba (@); en el caso del lenguaje XML, donde las diferentes partes se definen en forma de árbol, los delimitadores que señalan las etiquetas de los elementos son los caracteres "menor que" (<) y "mayor que" (>).

A continuación se puede apreciar un ejemplo de texto etiquetado en formato .xml:



Igualmente, *Connexor Machinese Syntax* nos da la opción de ver las relaciones sintácticas que ha establecido en imagen de árbol. Pinchando en cada una de las palabras podremos ver las etiquetas morfológicas de cada palabra. A continuación se muestra un ejemplo de árbol:

4. ETIQUETADO DE TXTCERAM Y TXTINFO-ES CON CONNEXOR

El objetivo de nuestro proyecto era trabajar con un etiquetador de textos automático para marcar dos corpus: TXTCeram y TXTInfo-ES. En el siguiente apartado se comentarán brevemente las pricipales características de cada corpus y se explicará el desarrollo del proceso de etiquetado con la herramienta Connexor Machinese Syntax.

4.1. Los corpus

4.1.1. El corpus TXTCeram

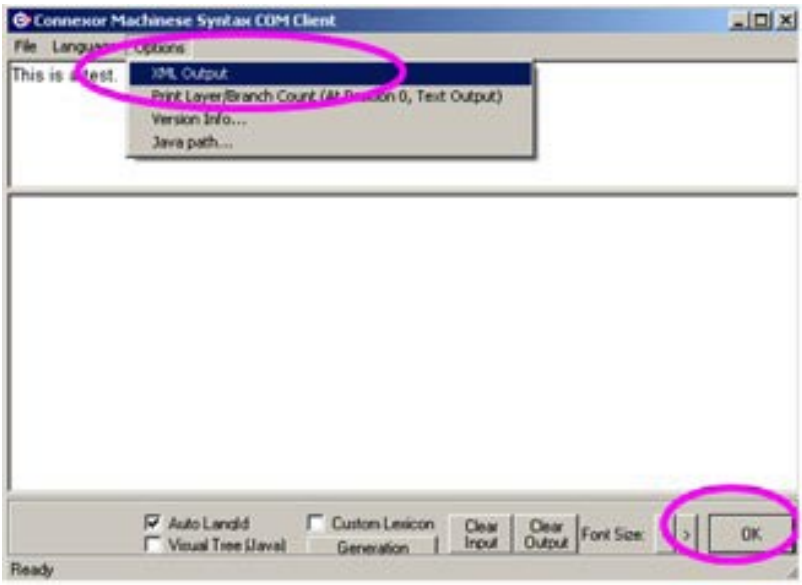
El corpus electrónico TXTCeram está compuesto por textos de especialidad del ámbito de la cerámica. Está formado por textos en español (alrededor de 2,8 millones de palabras) y textos en inglés (alrededor de 250.000 palabras). Uno de los objetivos con los partía el equipo del proyecto TXTCeram con la creación de este corpus fue el hecho de valorar la utilización de determinadas herramientas informáticas que pudieran resultar de utilidad, especialmente para los traductores, a la hora de hacer consultas, realizar extracciones terminológicas, analizar textos, etc.

4.1.1. El corpus TXTInfo-ES

El corpus TXTInfo-ES está formado por 580 textos en español (más de 505.000 palabras), en su mayoría didácticos, sobre Internet, la programación, la microinformática, la explotación de sistemas…. Se enmarca dentro del proyecto DiCoInfo del Observatoire de linguistique Sens-Texte (OLST) de la Université de Montréal (Canadá), dirigido por Marie-Claude L'Homme. La principal función de este corpus es la extracción terminológica para la creación de un diccionario básico de informática y de Internet.

4.2. El proceso

Etiquetar con el programa *Connexor Machinese Syntax* es una tarea sin muchas dificultades. En primer lugar, tenemos que seleccionar el archivo en formato .txt que queramos etiquetar, una vez aparezca el texto en la parte superior de la ventana configuraremos la opción de idioma. Si queremos etiquetar el texto en formato .xml seleccionaremos la opción "XML outpout". Para procesar el texto haremos clic en el botón "OK" que aparece en la parte inferior de la ventana, por último, sólo tendremos que guardar los resultados del marcado morfosintáctico del texto en el formato deseado.



4.2.1. Etiquetado de TXTCeram

Del corpus de TXTCeram, se han etiquetado solamente los textos en español. Para el corpus TXTCeram se etiquetaron 477 textos tanto en formato .txt como en formato .xml. El total de los textos tiene un tamaño de 14,3Mb. Cuando queremos etiquetar un corpus del tamaño de TXTCeram con el programa *Connexor Machinese Syntax,* el proceso, a pesar de ser sencillo, resulta demasiado mecánico, ya que al tratarse de tantos documentos y de la obtención de dos formatos diferentes tenemos que limpiar la ventana de resultados para seleccionar la opción "XML output" cada vez, y desmarcarla una vez realizado el análisis. Además, al ser los nombres de los documentos códigos numéricos, y por lo tanto similares entre ellos, hay que prestar atención al seleccionar los archivos para hacerlo en el orden adecuado sin olvidar ninguno por el camino.

4.2.2. Etiquetado de TXTInfo-ES

Para el corpus TXTInfo-ES se etiquetaron 580 textos también en ambos formatos, .txt y .xml. En este caso se trataba de un corpus un poco menos extenso que el de TXTCeram ya que tiene un tamaño de 3,07Mb. Cabe destacar, que los nombres de los textos que forman este corpus no eran códigos numéricos. Es por esto por lo que su identificación era más sencilla, puesto que se podría decir que

cada documento tiene un "nombre propio"; sin embargo, por este mismo motivo, el orden de los documentos era menos evidente (los archivos tienen nombres más complejos de identificar que un código numérico donde el último archivo siempre será el número más alto). Lo que también supuso un inconveniente a la hora de etiquetar el corpus, como en el caso del etiquetado de TXTCeram.

5. RESULTADOS DEL PROCESO

En general, los resultados obtenidos a partir del etiquetado con *Connexor Machinese Syntax* han sido tanto satisfactorios como útiles. Para valorar de una forma más exacta los resultados del etiquetado se decidió analizar el marcado morfosintáctico de una serie de segmentos seleccionados al azar. Se analizaron segmentos pertenecientes a tres textos del corpus TXTCeram y a tres textos del corpus TXTInfo-ES. Aún existiendo algunos fallos en el análisis gramatical de los elementos (comprendidos entre los fallos lingüísticos que se mencionan en el apartado 5.2), se puede concluir que los resultados obtenidos de dicho análisis son una muestra evidente de las ventajas del etiquetado con *Connexor*. Así, el total de fallos en el análisis de TXTCeram correspondía a un 6% de elementos etiquetados incorrectamente y el total de fallos encontrados en el análisis de los textos de TXTInfo-ES correspondía a un 10,7%. La diferencia entre el porcentaje de fallos puede deberse al tipo de textos que forman cada corpus. Los textos del corpus TXTCeram son en su mayoría artículos de investigación del ámbito de la cerámica de carácter divulgativo. En el caso del corpus de TXTInfo-ES, los textos son en general publicitarios y divulgativos y están escritos en un registro mucho más oral. Por lo tanto, los resultados del proceso son mejores cuando etiquetamos textos pertenecientes a un registro más formal y dirigidos a un público más especializado que cuando se trata de textos de contenido más general.

Sin embargo, se han presentado algunos inconvenientes que de ser subsanables, dotarían al programa de una mayor calidad. Los inconvenientes más destacados serán comentados a continuación.

5.1. Problemas técnicos

Así, el hecho de que el proceso se convirtiera en una actividad mecánica e incluso aburrida en ocasiones, se debe principalmente a que el programa no nos permitía analizar documentos .txt de un tamaño superior a los 36Kb. En este caso concretamente, estamos hablando del etiquetado de corpus que contienen textos bastante extensos, no tanto el corpus de TXTInfo-ES como el de TXTCeram. Por lo tanto, si el programa nos ofreciera la posibilidad de analizar los textos tal y como están divididos de forma natural, ganaríamos mucho más tiempo durante el proceso. De este modo, si no admite la carga de un archivo de tamaño superior a los 36Kb, nos vemos además obligados a realizar otra tarea extra: la de dividir los textos en textos más pequeños. Dicha tarea no puede realizarse automáticamente porque tenemos que asegurarnos de mantener una cierta coherencia respecto a los diferentes documentos para mantener las unidades de sentido y que no haya frases inacabadas al final de los mismos.

Otro de los inconvenientes de *Connexor Machinese Syntax* es que el programa nos da la opción de etiquetar el texto que hemos seleccionado como .xml mas en el momento de guardar los resultados no nos da la opción de "Guardar como tipo: Documento XML", tenemos que escribirlo manualmente; consideramos que ya que nos da la opción de etiquetar en XML sería más práctico guardar los resultados sin necesidad de escribir la extensión.

Por otro lado, en algunas ocasiones, en lugar de almacenar los documentos resultantes por orden alfabético el programa alteraba el orden de estos aleatoriamente, con lo cual, no nos podíamos guiar por el nombre del último archivo para saber cuál teníamos que analizar a continuación; esto obliga al usuario a tener que consultar qué archivo acaba de analizar para saber cuál va a continuación. En consecuencia, si el programa nos ofreciera la posibilidad de señalar de algún modo cuál ha sido el último texto etiquetado, dado el problema que supone el no poder etiquetar textos gran tamaño, el proceso se agilizaría notablemente y el usuario estaría más seguro de los pasos que realiza durante el proceso.

5.2. Problemas lingüísticos

En cuanto al análisis textual en sí mismo, en general *Connexor Machinese Syntax* etiqueta bien morfológicamente, pero hay que tener en cuenta que se pueden presentar ciertos fallos, a continuación se presentarán resumidamente los fallos más frecuentes que se han detectado:

relaciones sintácticas poco evidentes: cuando se trata de palabras con relaciones un poco más complicadas en las que los matices son difíciles de ver, incluso para los humanos, el ordenador etiqueta las palabras con la forma más evidente provocando que la relación no sea del todo exacta.
Ej. En la frase "Ideal para aquéllos que se inician en el mundo de la fotografía digital", Connexor ha marcado "ideal" como un sustantivo masculino singular cuando en realidad es un atributo.

palabras ambiguas: cuando una palabra puede tener dos funciones *Connexor Machinese Syntax* la etiqueta con ambas opciones, por lo que al no desambiguar la palabra no marca su relación sintáctica
Ej. En la frase "No siempre resulta sencillo llegar a la conclusión de que ha sufrido uno (ataque)", no desambigua la función sintáctica, etiqueta "uno" como pronombre masculino singular y como numeral cardinal masculino singular, y por lo tanto deja de establecer la relación sintáctica.

palabras en otros idiomas: si por casualidad *Connexor* encuentra algún extranjerismo, no lo detecta como tal y lo etiqueta como si se tratara de una palabra del lenguaje en el que está analizando
Ej. La construcción *on line*, *Connexor* marca "on" como un sustantitvo masculino singular y "line" como un sustantivo singular.

género no definido: cuando analiza palabras que al ser nombres propios, mayormente provinentes de otras lenguas en algunas ocasiones, en lugar de marcar un género ambiguo marca un género al azar

Ej. La palabra "Microsoft" está marcada por *Connexor* como un sustantivo femenino.

el pronombre reflexivo "se": en algunos casos el programa no establece correctamente la relación sintáctica entre el pronombre reflexivo y el verbo al que acompaña.
Ej. Verbo "recuperarse", *Connexor* etiqueta la partícula "se" de forma independiente y no siempre la relaciona con el verbo sino que a veces la relaciona con el objeto directo del verbo principal: "ayudar a las empresas a detectar y recuperarse de un ataque de red" (se >modifica ataque).

6. CONCLUSIONES

Se puede concluir que el etiquetado con *Connexor Machinese Syntax*, a pesar de los pequeños inconvenientes relativos al tamaño de los archivos y los pequeños fallos en el análisis, ha resultado ser un proceso sencillo y, tras el análisis realizado y desarrollado a lo largo del presente documento, con óptimos resultados a pesar de que los textos etiquetados aún no se hayan aplicado en los proyectos deseados.

Por otro lado, y haciendo referencia a los fallos en el análisis morfosintáctico, hay que considerar el tipo de textos por los que están formados los corpus. De este modo, el análisis es más efectivo, pues hay menos fallos en los resultados, cuando se trata de corpus con textos más técnicos como los de TXTCeram que cuando se trata de textos más divulgativos y más generales como los de TXTInfo-ES.

Asimismo, y respecto a los documentos resultantes del etiquetado, aún está en proceso la fase de aplicación, por lo que aún está por determinar si los resultados obtenidos cumplen todos los objetivos que requerimos para llevar a cabo nuestros proyectos de investigación.

Además, también está en proyecto un futuro estudio sobre la utilidad de los textos etiquetados producidos por *Connexor Machinese Syntax* en las extracciones terminológicas.

7. BIBLIOGRAFÍA

• Alcina Caudet, Mª Amparo (2005): *La implementación del concepto de género textual en los corpus electrónicos para traductores* en GARCÍA IZQUIERDO, ISABEL (ed.) El género textual y la traducción. Reflexiones teóricas y aplicaciones pedagógicas (Ed, García Izquierdo, I. c.) Peter Lang, Nueva York, pp. 93-114.

• Estellés Palanca, A., Mª Amparo Alcina Caudet y Victoria Soler Puertes (2006): *Retrieving Terminological Data from The TxtCeram Tagged Domain Corpus: First Step on a Terminological Ontology* en LREC 2006, Genova.

• Marie-Claude L'Homme (2007): *DiCoInfo Dictionnaire fondamental de l'informatique et de l'Internet*. Observatoire de linguistique Sens-Texte (OLST).

8. RECURSOS

Corpus:

• Institut Universitari de Lingüística Aplicada (IULA), Universitat Pompeu Fabra. [en línea] Bwananet, programa d'explotació del corpus tècnic de l'Iula.
http://bwananet.iula.upf.edu/bwananet1a.ca.htm [19/05/07]

• REAL ACADEMIA ESPAÑOLA: Banco de datos (CREA) [en línea]. Corpus de referencia del español actual. http://corpus.rae.es/creanet.html [19/05/07]

Etiquetadores:

• Connexor OY [en línea] http://www.connexor.com [19/05/07]

• Treetagger, Universidad de Stuttgart [en línea]
http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html [19/05/07]

• Herramienta de análisis morfológico Stilus, de Daedalus [en línea].
http://stilus.daedalus.es/herramientas.php?op=pos [19/05/07].

• Etiquetador morfológico de castellano del Grup d'Investigació en Lingüística Computacional (GILC), Universitat de Barcelona [en línea].
http://www.ub.es/gilcub/lascosas/eines/esdemos.html [19/05/07].