Desarrollo de un etiquetador morfosintáctico para el español

Fernando Sánchez León
Laboratorio de Lingüística Informática
Facultad de Filosofía y Letras
Universidad Autónoma de Madrid
fsanchez@ccuam3.sdi.uam.es

Amalio F. Nieto Serrano
Departamento de Ingeniería de Sistemas Telemáticos
Escuela Superior de Ingenieros de Telecomunicaciones
Universidad Politécnica de Madrid
anieto@dit.upm.es

12 de Junio de 1995

Abstract

This paper describes work performed withing the CRATER (*C*orpus *R*esources *A*nd *T*erminology *Ext**R**action*, MLAP-93/20) project, funded by the Commission of the European Communities. In particular, it addresses the issue of adapting the Xerox Tagger to Spanish in order to tag the Spanish version of the ITU (International Telecommunications Union) corpus. The model implemented by this tagger is briefly presented along with some modifications performed on it in order to use some parameters not probabilistically estimated. Initial decisions, like the tagset, the lexicon and the training corpus are also discussed. Finally, results are presented and the benefits of the *mixed model* justified.

Resumen

En este artículo se describe el trabajo realizado en el contexto del proyecto de investigación CRATER (*C*orpus *R*esources *A*nd *T*erminology *Ext**R**action*, MLAP-93/20), financiado por la Comisión de las Comunidades Europeas. En particular, se tratan los problemas de adaptación del Etiquetador Morfosintáctico de Xerox al español con el fin de etiquetar la versión española del corpus de la Unión Internacional de Telecomunicaciones (ITU). Se presenta brevemente el modelo implementado por este etiquetador junto con algunas modificaciones llevadas a cabo para incorporar en el sistema parámetros no estimados probabilísticamente. Asimismo, se discuten algunas decisiones iniciales, como el conjunto de etiquetas (*tagset*), el lexicón y el corpus de entrenamiento. Finalmente, se muestran los resultados y se justifican los beneficios de un *modelo mixto* como el propuesto.

Palabras clave: Etiquetado morfosintáctico, lingüística de corpus, modelos probabilísticos del lenguaje, estándares de etiquetado

1 Introducción

En este artículo se describe el trabajo de adaptación del Etiquetador Morfosintáctico de Xerox al español[1]. El Etiquetador de Xerox [Cutting *et al.*, 1992] presenta entre sus virtudes las de estar basado en un modelo probabilístico simple, como se verá a continuación, ser independiente de la lengua del corpus que se vaya a etiquetar y ser de dominio público[2]. Varios autores han realizado ya adaptaciones a otras lenguas distintas del inglés (lengua en la que lo han probado los autores). Así, [Feldweg, 1995] presenta los problemas de adaptación del etiquetador al alemán, mientras que [Chanod y Tapanainen, 1995] informan sobre el francés. Estas adaptaciones se han realizado de forma paralela a la que aquí se presenta para el español.

El interés por el Etiquetador de Xerox no sólo proviene de las virtudes antes mencionadas, sino que también se beneficia del atractivo que las aproximaciones estocásticas al Procesamiento del Lenguaje Natural han despertado de nuevo en los investigadores de este campo. Ejemplos ampliamente comentados de este resurgir de las técnicas probabilísticas incluyen el doble número especial que la revista *Computational Linguistics* dedicó recientemente a esta empresa. Más interesante, sin embargo, resulta, en este debate, la posibilidad de combinar técnicas (empíricas y racionalistas), en lugar de acercarse a los modelos estadísticos con la mentalidad de "a ver qué son capaces de hacer." Aunque lo cierto es que son capaces de "hacer cosas", con relativa sencillez en la estimación de parámetros y de una manera robusta, cualidad ésta que, ya es sabido, los sistemas basados en conocimiento no siempre poseen.

Fruto de esta combinación de técnicas, [Tapanainen y Voutilainen, 1994] presentan resultados inmejorables etiquetando corpus en inglés con el Etiquetador de Xerox y el sistema basado en restricciones ENGCG [Karlsson *et al.*, 1994]. En este caso, la combinación se realiza por medio de módulos separados. En el presente artículo, y de forma más modesta, se propone una combinación de técnicas dentro del propio Etiquetador de Xerox.

2 El Etiquetador Morfosintáctico de Xerox

El Etiquetador Morfosintáctico de Xerox utiliza un método estadístico para el etiquetado de texto. En estos sistemas, la ambigüedad de asignación de una etiqueta a una palabra se resuelve sobre la base de la interpretación más probable. Para ello, se utiliza una forma de modelo de Markov que supone que una palabra depende probabilísticamente sólo de su categoría gramatical que, a su vez, depende, en la mayoría de los sistemas aunque no en el Etiquetador de Xerox, solamente de la categoría de las dos palabras precedentes.

Dos son los tipos de entrenamiento utilizados con este modelo. El primero de ellos hace uso de un corpus de entrenamiento etiquetado. Se etiqueta manualmente una pequeña cantidad de texto que se utiliza para entrenar un modelo parcialmente ajustado. Este modelo se utiliza entonces para etiquetar más texto; las etiquetas se corrigen manualmente y el texto se utiliza posteriormente para volver a entrenar el modelo. Este método de entrenamiento se llama *bootstraping* [Derouault y Merialdo, 1986].

El segundo método no necesita un corpus de entrenamiento etiquetado. El modelo se denomina en este caso *modelo de Markov oculto* (*hidden Markov model, HMM*), pues las transiciones

[1]Este trabajo se ha realizado en el contexto del proyecto de investigación CRATER (*C*orpus *R*esources *A*nd *T*erminology *Ext**R**action*, MLAP-93/20), financiado por la Comisión de las Comunidades Europeas. En el proyecto participan, además de las instituciones a las que pertenecen los autores del presente artículo, la Universidad de Lancaster (Reino Unido), la empresa Computers, Communications and Visions, C^2V (Francia) e IBM-Francia.

[2]Puede obtenerse a través de `ftp` en `parcftp.xerox.com`, en el directorio `pub/tagger`. El programa se ejecuta bajo Common Lisp y se ha probado en varias implementaciones de este lenguaje para SunOS 4.x y 5.x, así como en la implementación para Macintosh.

entre estados no pueden determinarse mientras que se conoce la secuencia de elementos de salida. [Jelinek, 1985] utiliza este método para entrenar un etiquetador de texto. Generalmente, se emplea un modelo de trigramas en el que las estimaciones de trigramas se suavizan usando un método de interpolaciín lineal denominado *deleted interpolation* en el que las estimaciones de pesos se toman a partir de modelos de primer y segundo orden y de una distribución de probabilidades uniforme. [Kupiec, 1989] utiliza clases de equivalencia para palabras basadas en categorías, para reunir los datos de las palabras individuales. Las palabras más comunes se almacenan en un fichero de léxico, mientras que el resto de las palabras se representan según el conjunto de categorías posibles que pueden asumir. El número de clases de equivalencia (llamadas *clases de ambigüedad* en [Cutting *et al.*, 1992]) puede reducirse considerablemente[3]. Como reducción ulterior del número de parámetros, puede utilizarse un modelo de primer orden. En estos modelos, una palabra depende de su categoría gramatical, que a su vez depende solamente de la categoría de la palabra anterior.

El Etiquetador de Xerox está basado en un HMM. Utiliza clases de ambigüedad y un modelo de primer orden para reducir el número de parámetros que se han de estimar sin una reducción significativa de la precisión. De acuerdo con los autores, se pueden obtener resultados razonables entrenando el modelo con tan solo 3.000 oraciones. Además, "few ambiguity classes are sufficient for wide coverage, so it is unlikely that adding new words to the lexicon requires retraining, as their ambiguity classes are accomodated." Las palabras que no se encuentran en el diccionario reciben una clase de ambigüedad dependiendo tanto del contexto como de la información relacionada con el *sufijo* en el que terminan[4].

2.1 Procedimiento

El modo de operación del etiquetador es el siguiente. Después de que el *tokenizador* haya convertido el texto de entrada en una secuencia de *tokens*, estos tokens pasan al *lexicón*. Los tokens se convierten en un conjunto de *formas*[5], anotadas cada una de ellas con una etiqueta categorial. El conjunto de etiquetas identifica una *clase de ambigüedad*, que también devuelve el lexicón.

El módulo de *entrenamiento* toma grandes secuencias de clases de ambigüedad como entrada. Utiliza el algoritmo de Baum-Welch para producir un *HMM entrenado*, que constituye la entrada de módulo de etiquetado. El módulo de *etiquetado* recibe secuencias de clases de ambigüedad entre límites de oración. Estas secuencias se desambiguan computando el camino de máxima probabilidad a través del HMM con el algoritmo de Viterbi.

Las palabras que no se encuentren en el lexicón, que se ha de confeccionar a mano, "are generally both open class and regularly inflected", según [Cutting *et al.*, 1992][6]. Para calcular las clases de ambigüedad para estas palabras desconocidas puede emplearse un método específico para cada lengua. En este sentido, el Etiquetador de Xerox proporciona una función que computa los 'sufijos' junto con las predicciones probabilísticas sobre la categoría o categorías asignables a una palabra que termina en cada uno de los sufijos calculados. Esta función opera, asimismo,

[3] Aproximadamente 400 clases permiten cubrir el vocabulario completo del Brown Corpus. Este dato, sin embargo, depende no sólo de la lengua con la se esté trabajando, sino, principalmente, de la *finura* del conjunto de etiquetas utilizado.

[4] El término *sufijo* debe entenderse en este contexto en un sentido amplio (conjunto de caracteres finales en una palabra) y no estrictamente lingüístico.

[5] Los autores se refieren a estas formas como *stems*.

[6] Sin embargo, esto depende enormemente del tamaño del lexicón. Dado que los lexicones exhaustivos son "expensive, if not impossible to produce", en palabras de los autores, esta afirmación podría no ser cierta.

sobre un corpus de entrenamiento no etiquetado.

Como paso final, las palabras que no se encuentran en el diccionario y que terminan en un sufijo no reconocido reciben una clase de ambigüedad por defecto (la clase abierta).

3 Un modelo mixto

Como se ha mencionado en la sección anterior, cuando una palabra es desconocida para el sistema, puede utilizarse información relativa a su 'sufijo' con objeto de aproximar su posible clase de ambigüedad. Esta información puede calcularse por medio de la función LISP `class-guesser:train-guesser-on-files`. Los autores recomiendan firmemente el uso de esta función cuando se desee adaptar el etiquetador a nuevos corpus, nuevos conjuntos de etiquetas y nuevas lenguas [Jan Pedersen, comunicación personal]. Sin embargo, trataremos de demostrar que un sistema que utiliza un conjunto de sufijos incorporados manualmente arroja unos resultados mejores, al menos en una lengua flexiva como el español.

Esta función opera sobre un corpus de entrenamiento y calcula dos parámetros:

- los sufijos mismos
- la clase de ambigüedad asignada a cada sufijo

El único parámetro que puede controlarse en el cálculo de sufijos, es la longitud máxima. Esto puede realizarse cambiando el valor de la variable `class-guesser::*suffix-limit*`[7].

La clase de ambigüedad que se asignará a cada sufijo se selecciona del conjunto de clases computadas durante el entrenamiento normal, que se escribe en un fichero de clases. Este fichero contiene (i) todas y cada una de las etiquetas observadas en el lexicón (que son, obviamente, no ambigua), (ii) todos y cada uno de los conjuntos de etiquetas asignadas de forma ambigua para todas las formas del lexicón, y (iii) la clase de ambigüedad para la clase abierta (que es una clase por defecto).

La función anterior, tras computar un sufijo, observa las palabras del lexicón que terminan en el sufijo propuesto y el conjunto de etiquetas asignado a ellas. A continuación, elimina aquellas etiquetas que no estén incluidas en la clase de ambigüedad para la clase abierta y, entonces, trata de casar las clases restantes con una de las clases de ambigüedad existentes. Si tiene éxito, se asignará esta clase al sufijo. Por el contrario, si falla, el sufijo recibirá la clase de ambigüedad por defecto.

Mientras que este método puede ser correcto tanto para lenguas no flexivas (como el inglés) como para conjuntos de etiquetas reducidos, lo consideramos altamente ineficiente para lenguas flexivas y conjuntos de etiquetas más extensos. Trataremos de ejemplificar este extremo en el párrafo siguiente.

En el lexicón español existen muchas formas ambiguas. La mayoría de los casos oscila entre 2 a 4 etiquetas para cada forma, pero existen algunos casos que tienen incluso 5 o 6. Si establecemos, además, una clase abierta que incluya todas las etiquetas nominales (para sustantivos y adjetivos) y verbales, el fichero de clases contendrá, junto a esta clase abierta, una lista con todas las etiquetas individuales del *tagset*, varias clases de ambigüedad formadas por 2-tuplas, 3-tuplas, 4-tuplas y unas pocas 5-tuplas y 6-tuplas. Esto significa que los sufijos computados deben acomodarse en estas últimas clases de ambigüedad para maximizar la precisión en la

[7] Los autores han establecido este parámetro a 5.

asignación de etiquetas (el uso de la clase de ambigüedad por defecto en estos casos producirá resultados incorrectos en la mayoría de las ocasiones). Si suponemos que *a* es uno de los sufijos computados por la función, el problema entonces es tratar de casar el conjunto de etiquetas observadas en el lexicón para las palabras que terminan en *a* incluidas en la intersección con la clase de ambigüedad por defecto con una de las clases computadas previamente. Las palabras que terminan en *a* son, generalmente, adjetivos o nombres femeninos singulares y verbos en primera o tercera personas del singular del presente de subjuntivo o en tercera persona del singular del presente de indicativo (#(:ADJGFS :NCFS :VLPI3S :VLPS1S :VLPS3S))[8]. Ahora bien, si tomamos las formas del lexicón con 5 etiquetas diferentes, veremos que el número de clases generado por éstas se limita a cuatro:

```
#(:ADJGFS :ADJGMS :ADVGR :VLPPFS :VLPPMS)
#(:ADJGFS :ADJGMS :NCMS :VLPPFS :VLPPMS)
#(:ADJGFP :ADJGMP :NCMP :VLPPFP :VLPPMP)
#(:ADJGFS :ADJGMS :PREP :VLPPFS :VLPPMS)
```

Obviamente, no existe una correspondencia posible entre la primera clase de ambigüedad y ninguna de las segundas. Sencillamente, la primera clase de ambigüedad no existe: debe existir al menos una forma ambigua (que termine en *a* o en cualquier otro sufijo) que valide una clase de ambigüedad para que ésta pueda ser seleccionada cuando se calcule al observar las palabras que terminan en *a*. El resultado observado es que la función se ve obligada a asignar la clase de ambigüedad abierta a buena parte de los sufijos computados.

Asimismo, en las lenguas flexivas, la selección del corpus de entrenamiento es también crucial en la cuestión del cálculo de sufijos. Es preciso recoger y utilizar para el entrenamiento una cantidad suficiente de texto que contenga una variedad de palabras lo más amplia posible. Sin embargo, este requisito solo no garantiza una computación adecuada de los sufijos, dado que la función opera no sólo sobre los tokens del corpus de entrenamiento sino también sobre el lexicón del sistema. El parámetro que se ha de considerar en este sentido no es el tamaño real de este lexicón (que, en cualquier caso, es importante para asignar clases de ambigüedad a los tokens de un corpus con precisión), sino el conjunto de clases de ambigüedad representado en ese lexicón —y este conjunto no aumentaría con la incorporación de nuevas palabras.

Además, las lenguas flexivas, como el español, presentan la característica de tener una clara correspondencia entre los sufijos (lingüísticamente motivados) y las propiedades morfosintácticas de la palabra o palabras a las que se adjuntan. Consecuentemente, este conocimiento apriorístico podría explotarse en un sistema de etiquetado como el que aquí se describe. Por tanto, si una palabra que termina en *a* puede representar la siguiente clase de ambigüedad:

```
"a" #(:ADJGFS :NCFS :VLPI3S :VLPS1S :VLPS3S),
```

el sistema debería ser capaz de utilizar esta información sin necesidad de estimarla.

Por otra parte, la práctica de codificar manualmente la información para las palabras desconocidas se ha utilizado relativamente poco en los modelos probabilísticos del lenguaje. Algunos sistemas, como el Etiquetador de Xerox, computan probabilísticamente tanto los sufijos como las clases de ambigüedad asociadas a ellos; sin embargo, otros, como el que se descibe en [Weischedel *et al.*, 1993], incluyen una aproximación híbrida en la que los sufijos se añaden

[8] Algunos nombres y adjetivos masculinos pueden terminar en *a*. Algunas formas de imperativo también terminan en este sufijo. Sin embargo, estas palabras pueden tratarse como excepciones e incluirse en el lexicón.

a mano y las clases de ambigüedad se aproximan directamente a partir de los datos de entrenamiento.

Por otra parte, todos los etiquetadores probabilísticos utilizan información codificada a mano como, por ejemplo, un lexicón. Por tanto, una nueva aproximación podría incluir tanto tablas de sufijos como clases de ambigüedad codificadas a mano, especialmente para las lenguas flexivas en las que esta información puede obtenerse fácilmente. Esta modificación mejorará la precisión del sistema. Esta aproximación, sin embargo, presenta el inconveniente de que la migración del sistema a un nuevo conjunto de etiquetas implica un mayor esfuerzo de conversión de recursos, dado que tanto el lexicón como la tabla de sufijos tendrá que proyectarse a dicho conjunto de etiquetas.

A la luz de esta argumentación, se ha propuesto e implementado satisfactoriamente una modificación del sistema. Ésta consiste en combinar, durante la fase normal de entrenamiento, el conjunto de clases observado en el lexicón con las clases establecidas por un lingüista en el fichero de sufijos. El proceso de entrenamiento se beneficiará de la reducción del número de elementos de las clases de ambigüedad que se tienen que computar cuando se encuentran palabras no contenidas en el lexicón, mejorando la precisión en la generación de caminos de probabilidad.

En las siguientes secciones se muestran los beneficios de esta metodología de trabajo.

4 Ajuste del modelo

La estimación de parámetros es una cuestión fundamental en los modelos probabilísticos del lenguaje. Un modelo de Markov oculto del lenguaje puede ajustarse de distintas formas. Por tanto, se han tomado una serie de decisiones relativas al conjunto de etiquetas, al lexicón y a los "ajustes" (*biases*). Estas opciones se presentan y se justifican más abajo. También se comenta la selección de los datos de entrenamiento y los resultados obtenidos.

4.1 El conjunto de etiquetas y el lexicón

Los etiquetarios utilizados por los etiquetadores para el inglés generalmente se derivan, de alguna forma, del *tagset* utilizado en el Brown Corpus [Francis y Kučera, 1982], que distingue 87 etiquetas. La tendencia desde que se diseñó este etiquetario ha sido la de refinarlo y ampliarlo. Así, el Lancaster-Oslo/Bergen (LOB) Corpus distingue unas 135 etiquetas y el grupo de la UCREL, de la Universidad de Lancaster, utiliza un conjunto de 166 etiquetas (para CLAWS2 [Garside *et al.*, 1987]). Otros etiquetarios son incluso mayores, como el que utiliza el London-Lund Corpus of Spoken English, que contiene 197 etiquetas.

Estos refinamientos posteriores del *tagset* original del Brown Corpus reflejan la necesidad de que un corpus etiquetado muestre todas las idiosincrasias (morfo-)sintácticas de una lengua. Por tanto, el fundamento para desarrollar etiquetarios amplios y ricamente articulados es acercarse al "ideal of providing codings for all classes of words having distinct grammatical behaviour." [Sampson, 1987, 167]

Por otra parte, algunos proyectos basados en una orientación estocástica han modificado el *tagset* orginal del Brown Corpus reduciéndolo en lugar de ampliándolo. Éste es el caso del Penn Treebank Project, que utiliza 36 etiquetas categoriales [Marcus y Santorini, 1992]. Esta decisión se basó no sólo en el uso de un modelo probabilístico sino también en el hecho de que el objetivo era analizar sintácticamente el corpus, y que, por tanto, algunas distinciones categoriales eran

recuperables con referencia a la estructura sintáctica.

Sin embargo, las iniciativas internacionales encaminadas a la creación de estándares de anotación de corpus, como las propuestas por EAGLES [Leech y Wilson, 1994], recomiendan la distinción en los etiquetarios de las categorías morfosintácticas principales. De hecho, el nivel 1 (L1), que incluye *atributos/valores recomendados*, distingue, entre otras, *tipo*, *género*, *número*, *caso*, *person*, *tiempo*, *modo* y *forma personal/no personal*. Las recomendaciones de EAGLES manifiestan explícitamente que "[t]he standard requirement for these *recommended* attributes/values is that, if they occur in a particular language, then it is advisable that the tagset of that language should encode them." [Leech y Wilson, 1994, 16]

Consecuentemente, en la construcción de un etiquetario que se vaya a utilizar con un etiquetador probabilístico, debe encontrarse un equilibrio entre exhaustividad y precisión —cuanto más exhaustiva sea la información codificada en el etiquetario (cuanto mayor sea el etiquetario), menos preciso resultará el etiquetado (puesto que el modelo resultante será más complejo y la estimación de parámetros menos precisa).

Este equilibrio se ha tenido en cuenta en la creación del etiquetario para el español que se ha utilizado con el Etiquetador de Xerox en este proyecto. En un primer intento, se diseñó un etiquetario casi *ideal*, tomando en consideración no sólo las recomendaciones de EAGLES sino también las directrices de la TEI sobre anotación de textos [Simons, 1991], [Langendoen y Fahmy, 1991], [TEI AI1W2, 1991]. Este *etiquetario completo* se describe en [Sánchez-León, 1994]. Contiene 479 etiquetas (existen también etiquetas especiales para los signos de puntuación)[9]. Por tanto, se trata de un etiquetario muy amplio, que distingue casi todos los rasgos morfosintácticos recomendados por las iniciativas mencionadas. A continuación, se presentan algunos ejemplos de la información considerada para algunas categorías:

- Nombres: distinción *común/propio*, con varios subtipos para los propios; la información semántica que se tuvo en cuenta en el primer etiquetario (*temporal*, *locativo*, *medida*, *numeral* y *organización*) se ha restringido ahora sólo a *medida*, dada la gran cantidad de de trabajo de postedición derivado de la distinción inicial; otra información morfosintáctica habitual (*género*, *número*).

- Adverbios: *grado*, *carácter interrogativo*, *locativo* (con subtipos), *deixis* y *polaridad*.

- Verbos: *estatus* (léxicos/auxiliares), *persona*, *número*, *tiempo*, *modo*, *género* y *forma personal/no personal* (implícita). Dada la riqueza de la morfología verbal del español, las etiquetas verbales representan el 59% del número total de etiquetas.

Este etiquetario ha sido considerado "too finegrained to be suitable for a probabilistic tagger" [Lauri Karttunen, cominicación personal].

Por este motivo, se construyó un segundo *etiquetario reducido*, basado en el primero. El número de etiquetas se redujo drásticamente en este etiquetario a 174. Los rasgos considerados previamente para las categorías principales se han reducido a *género*, *número* y *persona*[10]. Las categorías menores también han visto reducida la información morfosintáctica (y a veces semántica) que se consideró en el primer etiquetario.

[9] La versión final que se utiliza actualmente es ligeramente distinta y contiene 466 etiquetas.

[10] Además, se ha tenido en cuenta el estatus de los verbos. Toda información semántica se ha eliminado en los sustantivos, si bien los nombres propios y los nombres de los días de la semana y de los meses poseen etiquetas específicas.

El etiquetario reducido se diseñó precisamente con la idea de comprobar la mejora en la precisión del etiquetado cuando el número de parámetros es menor.

Todos los etiquetadores probabilísticos hacen uso de un lexicón de distinto tamaño y cobertura. [Cutting *et al.*, 1992], por ejemplo, informan de los resultados de etiquetado de las oraciones pares del Brown Corpus usando un lexicón de 50.000 formas. Con este lexicón y el fichero de sufijos, no se encontró ninguna palabra desconocida durante el proceso de entrenamiento, por lo que no presentaron datos para el entrenamiento sobre formas a las que se les vaya a asignar la clase abierta.

Sin embargo, un lexicón mayor no garantiza necesariamente una mejor calidad del etiquetado. Las palabras son, por regla general, ambiguas y pueden recibir, dependiendo del contexto, una etiqueta categorial diferente. Sin embargo, la probabilidad de que una determinada palabra reciba una u otra categoría puede no ser la misma, por lo que algunos sistemas permiten especificar las posibles etiquetas asignables a una palabra ordenadas en una probabilidad descendente, e incluyen asimismo mecanismos especiales para expresar el hecho de que ciertas etiquetas son "raras" o "muy raras" [Garside *et al.*, 1987]. Cuando esta selección es imposible en el sistema, pueden emplearse otros recursos para reducir la ambigüedad. Algunos autores utilizan un lexicón óptimo que indica, para cada palabra, todas las etiquetas asignadas a ella en algún lugar del corpus que se esté utilizando, pero no otras etiquetas posibles aunque no recogidas en el corpus de trabajo [Merialdo, 1994]. Otros proponen la exclusión de lecturas raras del lexicón para evitar que el etiquetador las seleccione [Tapanainen y Voutilainen, 1994].

Como nuestro punto de partida no es un corpus etiquetado en el que realizar pruebas sobre un modelo estocástico determinado, el lexicón utilizado no está especialmente orientado hacia el corpus que pretendemos etiquetar. Por el contrario, nuestro deseo ha sido construir el etiquetador sobre un material léxico lo más uniforme posible. Por tanto, durante la construcción del mismo se ha tenido en cuenta el conjunto completo de etiquetas para cada palabra.

El lexicón utilizado por el sistema se ha generado compilando diferentes fuentes de información léxica, principalmente los diccionarios disponibles del proyecto de traducción automática EUROTRA [EUROTRA Dictionaries], aunque también se ha realizado codificación a mano. Este lexicón es el que se está utilizando en el etiquetado del corpus de ITU, dado que proporciona un modelo más preciso de la ambigüedad léxica que el proporcionado sólo por la información de sufijos[11].

4.2 Entrenamiento de un modelo de Markov oculto

El entrenamiento de los modelos de Markov ocultos se realiza sin un corpus etiquetado. En un etiquetador que funciona bajo este régimen, las transiciones entre estados (esto es, las transiciones entre categorías) no son observables. En estas circunstancias, el entrenamiento se realiza de acuerdo con el principio de Máxima Probabilidad (*Maximum Likelihood*), usando el algoritmo Forward-Backward (FB) o el de Baum-Welch. Este proceso de entrenamiento puede orientarse de diferentes formas para 'forzar' de alguna manera el proceso de aprendizaje. A continuación se describen dos de estas formas implementadas en el Etiquetador de Xerox, relacionadas con las clases de ambigüedad y con las transiciones entre estados:

[11] Sin embargo, este lexicón ha de manejarse con cuidado. El diccionario bilingüe de Collins no está libre de errores. De hecho, la información morfosintáctica de la parte española es a menudo incorrecta. Se ha corregido sistemáticamente sólo para los paradigmas verbales, pero no para los paradigmas nominales ni para la asignación categorial. Esta revisión exhaustiva se está realizando actualmente.

- Los ajustes sobre clases de ambigüedad se llaman *ajustes de símbolos* (*symbol biases*). Éstos representan un tipo de probabilidad léxica para determinadas clases de ambigüedad. De este modo, las clases de ambigüedad se anotan con etiquetas favorecidas. Obsérvese, sin embargo, que esta preferencia se establece para una clase determinada y no para las formas individuales del lexicón (como se hace, por ejemplo, en CLAWS [Garside *et al.*, 1987]), por lo que el mecanismo es menos eficiente que en otros sistemas.

- Los ajustes sobre transiciones entre estados se llaman *ajustes de transiciones* (*transition biases*). Éstos especifican la probabilidad o no de que a una etiqueta le sigan una o varias etiquetas específicas. Los ajustes pueden formularse bien como probabilidades favorecidas bien como desfavorecidas. Las probabilidades desfavorecidas reciben una pequeña constante pero no se prohíben; por el contrario, los datos del corpus de entrenamiento pueden modificar estas probabilidades iniciales.

En este sentido, el modelo ha sido inicialmente ajustado usando tanto ajustes de transiciones como de símbolos. El número de ajustes de transiciones utilizados es de tan sólo 5 y el de ajustes de símbolos 3[12]. Los del primer tipo incluyen el favorecimiento de las transiciones clítico-verbo, determinante-nombre y nombre-adjetivo, y el desfavorecimiento de las transiciones adjetivo-adjetivo y preposición-verbo en forma personal. Veamos un ejemplo (ligeramente simplificado):

```
(:valid :ppo3fs
        :veci3p :veci3s :vefi3p :vefi3s :vefs3p :vefs3s :veii3p :veii3s
        :veis3p :veis3s :vepi3p :vepi3s :veps3p :veps3s :vexi3p :vexi3s
        :vhci3p :vhci3s :vhfi3p :vhfi3s :vhfs3p :vhfs3s :vhii3p :vhii3s
        :vhis3p :vhis3s :vhpi3e :vhpi3p :vhpi3s :vhps3p :vhps3s :vhxi3p
        :vhxi3s :vlci3p :vlci3s :vlfi3p :vlfi3s :vlfs3p :vlfs3s :vlii3p
        :vlii3s :vlis3p :vlis3s :vlpi3p :vlpi3s :vlps3p :vlps3s :vlxi3p
        :vlxi3s :vmci3p :vmci3s :vmfi3p :vmfi3s :vmfs3p :vmfs3s :vmii3p
        :vmii3s :vmis3p :vmis3s :vmpi3p :vmpi3s :vmps3p :vmps3s :vmxi3p
        :vmxi3s)
```

Este ajuste favorece la transición entre un pronombre proclítico de tercera persona de singular femenino en función de objeto directo (*la*, en *la amó*) y una forma verbal finita de cualquier tipo de verbo (excepto *ser*)[13] en cualquier combinación de tiempo, modo, persona y número (excepto imperativo)[14].

Los ajustes de símbolos incluyen el favorecimiento de las lecturas de conjunción para *y*, *e*, *o*, *u* y de preposición para *a* frente a la posible lectura como letra del alfabeto y el de la etiqueta nominal para aquellas palabras que también pueden ser adjetivos[15]. Así, una palabra que pueda ser nombre o adjetivo masculino singular recibirá preferentemente la etiqueta nominal:

```
(:valid (:adjgms :ncms) :ncms)
```

[12] El número real de ajustes es mayor porque es necesario escribir un ajuste para cada etiqueta que designa un determinante o un clítico (con distinción de género, número, caso, etc.) y las etiquetas a su derecha.

[13] Las etiquetas que comienzan con ve son para el verbo *estar*, vh designa las del verbos *haber*, los verbos con contenido léxico llevan una etiqueta que comienza por vl, vm identifica a los modales.

[14] El ajuste sólo muestra las formas de tercera persona por simplificación.

[15] Las caraterísticas del corpus de ITU, formado por textos de telecomunicaciones, han motivado esta decisión.

[Tapanainen y Voutilainen, 1994], que, como se ha dicho, utilizan el Etiquetador de Xerox en combinación con ENGCG para etiquetar textos en inglés, con un porcentaje de acierto del 98,5%, proponen otros medios para ajustar el sistema. Éstos son los siguientes:

- No incluir las lecturas raras en el lexicón para evitar que el etiquetador las seleccione.
- Utilizar valores diferentes para el número de iteraciones (el número de veces que se utiliza el mismo bloque de texto en el entrenamiento) y para el tamaño del bloque de texto que se utiliza para entrenar.
- La selección del corpus de entrenamiento afecta también al resultado final.

En nuestro caso, ya se ha comentado que se tomó la decisión previa de probar el sistema sin limitaciones léxicas especiales, esto es, con un lexicón que reflejara el conjunto completo de etiquetas asignables a cada palabra en él contenida. Con respecto a la segunda sugerencia, se han conservado los parámetros iniciales propuestos por los desarrolladores del Etiquetador de Xerox con el fin de no introducir mayor complejidad en la estimación inicial de parámetros. Finalmente, la elección del corpus de entrenamiento tiene consecuencias en la precisión del sistema. Como ha demostrado [Merialdo, 1994], cuando se utiliza un HMM, un corpus de entrenamiento mayor no necesariamente garantiza una mayor precisión. Por el contrario, un modelo inicial estimado realizando un entrenamiento basado en Frecuencia Relativa (*Relative Frequency, RF*) sobre un texto etiquetado puede degradarse si a continuación se utiliza como corpus de entrenamiento un corpus no etiquetado relativamente grande. En nuestro caso, no ha sido posible realizar un entrenamiento combinado (RF y ML) pero, en cualquier caso, se ha tenido en cuenta la degradación potencial del modelo cuando se ha producido el modelo final.

4.3 Entrenamiento y resultados

Se ha entrenado el sistema utilizando ambas versiones del etiquetario. Aunque ya se había tomado la decisión de etiquetar el corpus de ITU usando la versión *completa* del etiquetario y, de hecho, ya ha comenzado la postedición del corpus etiquetado con ésta, se ha desarrollado de forma paralela una versión del etiquetador que utiliza el *etiquetario reducido*. Los resultados obtenidos con ambos conjuntos de etiquetas se presentan en esta sección.

Se ha utilizado como corpus de entrenamiento el subcorpus de 1 millón de palabras que se va a posteditar, dejando el fichero `SP_itu_corpus_000` como corpus de prueba. Este corpus contiene 9.366 tokens. El corpus de entrenamiento se ha utilizado de forma incremental, comprobando los resultados con cada modelo parcial obtenido.

Con ambos etiquetarios, el sistema utilizado incluye un conjunto inicial de ajustes de transiciones y de símbolos que es responsable de los buenos resultados obtenidos con el modelo uniforme (sin entrenamiento). Los ajustes son los mismos para cada modelo, así como el lexicón (en términos de cobertura) y el fichero de información de sufijos.

Como podrá comprobarse cuando se observen los resultados, no existe una curva de aprendizaje clara. El sistema se comporta de forma relativamente aceptable con el conjunto de ajustes iniciales y su precisión mejora incluso un 2,5% con una pequeña cantidad de texto. Sin embargo, los mejores resultados se obtienen sólo 50.000 palabras, siendo la precisión obtenida con corpus mayores casi la misma. En cualquier caso, puesto que los resultados obtenidos con los otros

modelos son tan parecidos, resulta difícil probar la afirmación de Merialdo.

Con respecto a la comparación entre ambos etiquetarios, la curva es la misma en los dos casos, habiéndose obtenido también los mejores resultados con la misma cantidad de texto de entrenamiento. Sorprendentemente, la precisión es también la misma para ambos etiquetarios con el modelo mejor. Sin embargo, en general, el *etiquetario reducido* muestra un insignificante 0,1% de mejora en la precisión con respecto al *etiquetario completo*[16].

La tabla 1 muestra el comportamiento del sistema cuando se etiqueta el corpus de prueba con el etiquetario *completo*.

Tabla 1: Resultados obtenidos con el etiquetario *completo*.

Ficheros de entrenamiento	Recuento de palabras[a]	Tiempo de aprendizaje[b]	Errores etiquetando el corpus de prueba[c]	Precisión
Sin entrenamiento	0	–	1059 - 645	88.69 - 93.11
001-003	29931	30'24"	819 - 405	91.26 - 95.68
001-006	53300	53'55"	790 - 376	91.51 - 95.93
001-009	66922	1h 06'48"	855 - 441	90.87 - 95.29
001-014	96603	1h 37'01"	840 - 426	91.03 - 95.45
001-019	143129	2h 23'14"	853 - 439	90.89 - 95.31
001-024	180302	2h 59'47"	830 - 416	91.14 - 95.56
001-029	213518	3h 27'55"	827 - 413	91.17 - 95.59
001-034	255960	4h 21'03"	835 - 421	91.08 - 95.50
001-039	293203	4h 52'52"	833 - 419	91.11 - 95.53
001-044	333570	5h 26'52"	832 - 418	91.12 - 95.54
001-049	371338	6h 05'20"	835 - 421	91.08 - 95.50
001-054	401433	6h 38'39"	833 - 419	91.11 - 95.53
001-059	424189	6h 58'21"	832 - 418	91.12 - 95.54
001-064	427487	Sin memoria	–	–
001-069	507608	8h 16'51"	829 - 415	91.14 - 95.56
001-074	586608	9h 36'16"	828 - 414	91.16 - 95.58
001-079	637565	10h 15'57"	835 - 421	91.08 - 95.50
001-084	698788	11h 13'50"	834 - 420	91.10 - 95.52
001-089	776407	12h 25'55"	829 - 415	91.14 - 95.56
001-094	823498	13h 20'34"	827 - 413	91.17 - 95.59
001-099	890247	14h 16'14"	825 - 411	91.19 - 95.61
001-108	971163	15h 47'20"	832 - 418	91.12 - 95.54

[a] Recuento realizado con el comando wc de UNIX

[b] Tiempo real

[c] El primer dato representa el número absoluto de errores; el segundo no incluye las palabras extranjeras

La tabla 2 muestra el comportamiento del sistema al etiquetar el corpus de prueba con el etiquetario *reducido*.

5 Un modelo enriquecido lingüísticamente

Además de las razones mencionadas en las secciones anteriores relativas a la adecuación de un modelo basado en el conocimiento lingüístico para tratar la información de sufijos, al menos en las lenguas flexivas, existe una razón de carácter pragmático: el etiquetado debería ser más preciso utilizando un modelo enriquecido lingüísticamente que con el original, sólo estadístico. Con objeto de probar esta afirmación, se va a realizar una comparación de los resultados con

[16] Todos los resultados presentados se refieren a la versión 1.2 del Etiquetador de Xerox. Con la versión 1.1, el entrenamiento producido es siempre el mismo independientemente del corpus de entrenamiento usado. No es sorprendente, por tanto, que el fichero HMM sea también el mismo y que los tiempos de entrenamiento sean sospechosamente cortos. Puede concluirse que la versión 1.1 parece no aprender nada del corpus de entrenamiento.

Tabla 2: Resultados obtenidos con el etiquetario *reducido*.

Ficheros de entrenamiento	Recuento de palabras[a]	Tiempo de aprendizaje[b]	Errores etiquetando el corpus de prueba[c]	Precisión
Sin entrenamiento	0	-	1032 - 618	88.98 - 93.40
001-003	29931	20'12"	804 - 390	91.42 - 95.84
001-006	53300	37'02"	790 - 376	91.51 - 95.93
001-009	66922	46'00"	848 - 434	90.95 - 95.37
001-014	96603	1h 09'43"	836 - 422	91.07 - 95.49
001-019	143129	1h 40'30"	827 - 413	91.17 - 95.59
001-024	180302	2h 05'14"	825 - 411	91.19 - 95.61
001-029	213518	2h 25'07"	819 - 405	91.26 - 95.68
001-034	255960	2h 56'57"	822 - 408	91.22 - 95.64
001-039	293203	3h 19'50"	837 - 423	91.06 - 95.48
001-044	333570	3h 47'51"	832 - 418	91.12 - 95.54
001-049	371338	4h 20'24"	831 - 417	91.13 - 95.55
001-054	401433	4h 35'15"	823 - 409	91.21 - 95.63
001-059	424189	4h 48'16"	820 - 406	91.24 - 95.66
001-064	427487	4h 51'21"	820 - 406	91.24 - 95.66
001-069	507608	5h 45'26"	818 - 404	91.27 - 95.69
001-074	586608	6h 41'45"	818 - 404	91.27 - 95.69
001-079	637565	7h 12'32"	818 - 404	91.27 - 95.69
001-084	698788	7h 50'53"	813 - 399	91.32 - 95.74
001-089	776407	8h 40'21"	816 - 402	91.29 - 95.71
001-094	823498	9h 13'37"	812 - 398	91.33 - 95.75
001-099	890247	9h 54'11"	817 - 403	91.28 - 95.70
001-106	971163	10h 54'24"	821 - 407	91.23 - 95.65

[a]Recuento realizado con el comando wc de UNIX

[b]Tiempo real

[c]El primer dato representa el número absoluto de errores; el segundo no incluye las palabras extranjeras

ambos modelos. Por el momento, se pueden comparar los ficheros de sufijos obtenidos con los dos métodos con objeto de determinar *a priori* cuál pueda ser el mejor.

Para la función que computa automáticamente la información sobre sufijos, se ha utilizado el subcorpus completo que se va a posteditar. Los resultados obtenidos tanto a mano como automáticamente se presentan en la tabla 3:

Tabla 3: Información del fichero de sufijos.

Modelo	Incorporados manualmente		Computados automáticamente							
Etiquetario	completo	reducido	completo				reducido			
Modelo entrenado previamente	-	-	no		sí		no		sí	
Parámetro de longitud máxima de sufijos	-	-	15	5	15	5	15	5	15	5
Número de sufijos	208[a]	208[b]	16	94	16	100	16	78	16	77
Longitud máxima de sufijos	-	-	1	4	1	4	1	4	2	4
Número total de etiquetas	376	362	97	445	97	340	87	418	51	311
Etiquetas por sufijo	1.8	1.7	6	4.7	6	3.4	5.4	5.4	3.2	4.1

[a]Además, este fichero incluye 306 sufijos para el reconocimiento de verbos con enclíticos y 22 sufijos para palabras extranjeras.

[b]Además, este fichero incluye 306 sufijos para el reconocimiento de verbos con enclíticos y 22 sufijos para palabras extranjeras.

Obsérvese que la función que calcula automáticamente la información relativa a los sufijos puede ejecutarse con un modelo entrenado o no entrenado. Los resultados, sin embargo, son mejores con un modelo entrenado previamente. Sin embargo, estos resultados están lejos de los obtenidos cuando la información se incluye manualmente. Además, el número de sufijos es menor y la longitud máxima no garantiza el reconocimiento de sufijos típicamente no ambiguos: *-mente*, que es siempre un adverbio, o *-ción*, que siempre es un sustantivo femenino singular[17]. Otra desventaja importante de la función es que no tiene en cuenta la caja de los caracteres,

[17]El corpus que se está etiquetando se ha convertido a una representación SGML plana, especialmente en lo relativo a los caracteres de 8 bits. Obsérvese que la representación SGML de los caracteres ISO LATIN convierte el último sufijo en *-ción*, por lo que resulta imposible su identificación con un límite de sufijo de 5 caracteres.

produciendo por tanto sufijos tanto en mayúsculas como en minúsculas con información dif rente en cada caso.

Consecuentemente, las prestaciones de un modelo que utilice la aproximación propuesta se mejor para el español que las de la estrategia original del etiquetador.

6 Otras cuestiones

El Etiquetador de Xerox carece de los mecanismos adecuados para tratar los elementos léxicc complejos. La segmentación del texto en tokens se realiza a través de la información gráfica com caracteres de espaciado y otros delimitadores. Esto plantea un problema para la identificació tanto de palabras ortográficas complejas que comprenden más de una palabra textual (po ejemplo, formas verbales con enclíticos) como de palabras textuales que abarcan más de un palabra ortográfica (por ejemplo, las unidades multipalabra continuas invariantes).

La primera cuestión incluye la segmentación de palabras de orden superior para el re conocimiento y posterior etiquetado de las formas verbales con enclíticos. En una primer versión, el sistema asignaba una etiqueta especial, `VCLI`, a estas formas. Sin embargo, se ha introducido rutinas especiales para la segmentación inteligente (comprobando la correcta orde nación de lo clíticos) de las formas verbales con enclíticos, de forma que el sistema etiquet correctamente estos elementos. Sin embargo, la complejidad de esta tarea pone de manifiest una de las grandes limitaciones del Etiquetador de Xerox: el sistema carece de un analizado morfológico. Esta limitación pone en tela de juicio una de las afirmaciones de los desarrolladore del sistema: su independencia de la lengua de etiquetado. El repertorio léxico de formas flexionadas de lenguas con una alta productividad flexiva puede saturar el sistema. Otro tanto ocurre con lenguas altamente aglutinantes, en las que los mecanismos de formación de palabras hacen imposible la creación de un lexicón medianamente amplio.

La segunda cuestión se ha resuelto por medio de una fase de preprocesamiento. Los caracteres de espacio que separan los componentes de una palabra textual compleja se reemplazan por una tilde (~), con lo que la tokenización normal puede operar sobre ellos. Para este fin, se ha utilizado un programa desarrollado por Theo W. Tams, de EUROTRA-DK [Jensen *et al.*, 1990], durante la tercera fase de EUROTRA-I para una propuesta sobre Integración de Interfaces Finales. El código ha sido adaptado a nuestros requisitos.

Junto a éstas, se han llevado a cabo algunas otras modificaciones sobre el Etiquetador de Xerox original. Así, la sintaxis de salida de las etiquetas se ha modificado, de forma que en lugar de presentarse en la línea inferior se muestran a la derecha de la palabra separadas de ésta por un signo de subrayado (_), como es tradicional en los etiquetadores de la literatura anglosajona, especialmente en los trabajos de la Universidad de Lancaster.

El tokenizador ha sido adaptado para que pueda tratar entidades SGML dentro de las palabras así como para reconocer lindes de oración cuando en posición inicial de la oración siguiente se encuentra una entidad SGML o un signo de abertura de exclamación o admiración.

El tokenizador identifica correctamente los lindes oracionales, con las usuales limitaciones inherentes a esta tarea, como, por ejemplo, la distinción adecuada entre los puntos de final de oración y los de las abreviaturas. Se trata de una cuestión primordial para el comportamiento del sistema dado que el proceso de entrenamiento se realiza sobre fragmentos de texto segmentados en oraciones.

Sin embargo, dado que el punto final se utiliza para identificar oraciones, no puede ser etiquetado por el propio tokenizador. Una fase de postedición automática, que realiza actualmente

también la corrección de algunas transiciones entre categorías que el sistema tiende a etiquetar incorrectamente, se encarga de resolver este problema.

También se han implementado algunas reglas especiales de tokenización para reconocer dos formatos de fecha que aparecen en el corpus de ITU: `dd.mm.aa` y `aaaa-aaaa`.

7 Conclusiones

En este artículo, se presentan los resultados obtenidos con la adaptación para el español de un etiquetador de dominio público, el Etiquetador de Xerox. Con algunas modificaciones, necesarias a nuestro juicio, para el etiquetado de lenguas flexivas (tratamiento de la información sobre sufijos) y para la adecuada segmentación de palabras complejas (formas verbales con enclíticos), el sistema no sobrepasa los porcentajes de error aceptados para otros etiquetadores morfosintácticos. El modelo, de momento, no se ha probado con texto libre, sino sólo con el corpus de ITU. Los resultados pudieran ser inferiores en este caso. [Chanod y Tapanainen, 1995] presentan, utilizando un corpus distinto al corpus de entrenamiento, resultados sensiblemente superiores para el francés a los obtenidos por nosotros para el español (96,8% de precisión), mientras que la precisión para el alemán se cifra en 96,66% [Feldweg, 1995]. Sin embargo, debe tenerse en cuenta la diferencia entre los etiquetarios utilizados por estos autores y el etiquetario utilizado en este proyecto. Mientras que [Chanod y Tapanainen, 1995] utilizan un conjunto con 88 etiquetas y [Feldweg, 1995] uno con 42, nuestro etiquetario tiene 466 etiquetas diferentes. Este número de etiquetas puede ser excesivo, especialmente para un etiquetador probabilístico, pero los resultados obtenidos con el corpus de trabajo demuestran una precisión similar, con el valor añadido para el corpus etiquetado de presentar toda la variedad de categorías y subcategorías morfosintácticas reflejada en el mismo[18].

Agradecimientos
Queremos agradecer a Flora Ramírez Bustamante sus comentarios y ayuda en la construcción de los recursos lingüísticos sobre los que se ha desarrollado este etiquetador. Ruthanna Barnett también ha aportado su granito de arena con sus comentarios.

Referencias

[Nieto, 1994] A. F. Nieto. *CRATER: UPM Progress for the Period April-September 1994*. CRATER Internal Document. September 1994.

[COLLINS Spanish-English, 1994] *Collins Concise Spanish/English Bilingual Dictionary*. Electronic Edition.

[Cutting *et al.*, 1992] D. Cutting, J. Kupiec, J. Pedersen, y P. Sibun. A Practical Part-of-Speech Tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento.

[Chanod y Tapanainen, 1995] J.-P. Chanod y P. Tapanainen.Tagging French – comparing a statistical and a constraint-based method. In *Proceedings of the EACL-95*, Dublin.

[Derouault y Merialdo, 1986] A. M. Derouault and B. Merialdo. Natural Language Modelling for Phoneme-to-Text Transcription. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8:742–749.

[EUROTRA Dictionaries] EUROTRA Spanish Monolingual Dictionary. 1992.

[Feldweg, 1995] H. Feldweg. Implementation and evaluation of a German HMM for POS disambiguation. In *Proceedings of the EACL SIGDAT workshop 1995*, Dublin.

[18] La versión española de este etiquetador estará en el dominio público en el mes de octubre de 1995.

[Francis y Kučera, 1982] W. N. Francis y H. Kučera. *Frequency analysis of English usage. Lexicon and grammar*, Houghton Mifflin, Boston.

[Garside *et al.*, 1987] R. Garside, G. Leech and G. Sampson. *The Computational Analysis of English. A Corpus-Based Approach*, Longman, London.

[Jelinek, 1985] F. Jelinek. Markov Source Modeling of Text Generation. In J. K. Skwirzinski, editor, *Impact of Processing Techniques on Communication*, Nijhoff, Dordrecht.

[Jensen *et al.*, 1990] N. Jensen, T. Tams, N. Jaeger y V. Pirrelli. *Final Report on Front End Integration*, EUROTRA Internal Document.

[Karlsson *et al.*, 1994] F. Karlsson, A. Voutilainen, J. Heikkilä y A. Anttila (eds.) *Constraint Grammar: a Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin.

[Kupiec, 1989] J. M. Kupiec. Probabilistic Models of Short and Long Distance Word Dependencies in Running Texts. In *Proceedings of the 1989 DARPA Speech and Natural Language Workshop*, pages 290–295, Morgan Kaufman, Philadelphia.

[Langendoen y Fahmy, 1991] D. T. Langendoen y E. Fahmy. *Feature-structure markup for presentation at Oxford and Brown workshops*, Department of Linguistics, University of Arizona, Tucson, AZ 85721 USA, September.

[Leech y Wilson, 1994] G. Leech y A. Wilson. *Draft Sections 4.6 and 4.7 ofthe EAGLES Interim Report: Annotation Sub-Group*, EAGLES, February.

[Marcus y Santorini, 1992] M. P. Marcus y B. Santorini. *Building very large natural language corpora: the Penn Treebank*, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104, January.

[Merialdo, 1994] B. Merialdo. Tagging English Text with a Probabilistic Model. *Computational Linguistics*, 20(2), 155–171.

[Sampson, 1987] G. Sampson. Alternative grammatical coding systems. In Garside *et al.*. *The Computational Analysis of English. A Corpus-Based Approach*, Longman, London, 165–183.

[Sánchez-León, 1994] F. Sánchez León. *Spanish tagset for the CRATER project*, CRATER Internal Document, March. También disponible a través de WWW como http://xxx.lanl.gov/cmp-lg/9406023.

[Simons, 1991] G. F. Simons. *Feature System Declarations and the Interpretation of Feature Structures*, January 1991.

[Tapanainen y Voutilainen, 1994] P. Tapanainen and A. Voutilainen. Tagging accurately – Don't guess if you know. To appear in *Proceedings of the Fourth Conference on Applied Natural Language Processing*, Stuttgart.

[TEI AI1W2, 1991] Text Encoding Initiative. *TEI AI 1W2. List of Common Morphological Features For Inclusion in TEI Starter Set Of Grammatical-Annotation Tags*, June.

[Weischedel *et al.*, 1993] R. Weischedel, M. Meteer, R. Schwartz, L. Ramshaw, y J. Palmucci. Coping with Ambiguity and Unknown Words through Probabilistic Models. *Computational Linguistics*, 19(2), 359–382.