Análisis morfosintáctico estadístico en lengua gallega *

Francisco Méndez Pazó
Universidad de Vigo
E.T.S.I. Telecomunicación
fmendez@gts.tsc.uvigo.es

Francisco Campillo Díaz
Universidad de Vigo
E.T.S.I. Telecomunicación
campillo@gts.tsc.uvigo.es

Eduardo Rodríguez Banga
Universidad de Vigo
E.T.S.I. Telecomunicación
erbanga@gts.tsc.uvigo.es

Elisa Fernández Rei
Universidad de Santiago
Facultad de Filología
fgelisa@usc.es

Resumen: En este artículo describimos la construcción de un analizador morfosintáctico en gallego que, además de su evidente interés lingüístico, sea fácilmente aplicable a sistemas de reconocimiento y síntesis de voz. Los modelos estadísticos han demostrado que son capaces de ofrecer unas prestaciones similares a sistemas que emplean innumerables reglas intrincadas que, por otro lado, son muy difíciles de depurar y mantener. Por el contrario los modelos estocásticos permiten un diseño rápido, si se dispone de un corpus de entrenamiento, y son extremadamente flexibles, ya que pueden ser adaptados a otro idioma sin modificaciones excesivas del código. Para entrenar los modelos estadísticos se ha comenzado la recogida de un corpus en gallego que, por el momento, consta de unas 400.000 palabras etiquetadas morfosintácticamente.
Palabras clave: Análisis morfosintáctico, análisis estadístico, corpus gallego

Abstract: This paper describes a morphosyntactic analyser in Galician which, apart from its obvious linguistic interest, can be easily applied to speech recognition and speech synthesis systems. While rule-driven models produce the better performance, stochastic models have shown a comparable accuracy when properly designed. Moreover, rule-driven models are based on a complex set of linguistic rules, quite difficult to maintain and not directly extensible to other languages. On the contrary, stochastic models allow a quick design, if a training corpus is available, and are extremely flexible as they can be adapted to other languages with minor changes in their source code. In order to train the statistic models we began to collect a Galician corpus which, at this time, consists of about 400,000 words with morphosyntactic annotations.
Keywords: POS tagging, morphosyntactic analysis, Galician corpus

1. Introducción

Hoy en día, el gran avance experimentado por las denominadas Tecnologías del Habla hace necesario el desarrollo de técnicas de análisis morfosintáctico precisas y flexibles. El análisis morfosintáctico es una etapa clave de cualquier sistema de conversión texto-voz y, por otra parte, su incorporación a los sistemas de reconocimiento de habla continua aumenta significativamente las tasas de reconocimiento.

A la hora de abordar la construcción de un analizador morfosintáctico, hay básicamente dos alternativas: la primera, puramente lingüística, se basa en la obtención de un conjunto de reglas por parte de expertos lingüistas; la segunda, construye un modelo estadístico del lenguaje en el que cada secuencia de categorías morfosintácticas tiene asociada una probabilidad de aparición, que es estimada a partir de una cantidad relativamente grande de texto correctamente etiquetado((Brants, 2000) (Brill, 1995)(Cutting et al., 1992)(Ratnaparkhi, 1996)(Daelemans et al., 1996)). Obviamente, el diseño de un sistema preciso totalmente basado en reglas es extremadamente complicado debido no sólo a la necesidad de expertos, a la falta de información semántica y al gran número de reglas que

* Este trabajo ha sido parcialmente financiado por el Ministerio de Ciencia y Tecnología, fondos Feder y la Xunta de Galicia, en los proyectos TIC2002-02208, PGIDT01PXI32205PN y PGIDT02PXI32201PR.

son necesarias, sino también a la adecuada secuenciación de las reglas, ya que su correcto funcionamiento depende del orden en que se apliquen. Por contra, los métodos estadísticos proporcionan una forma más sencilla y flexible de desarrollar un analizador morfosintáctico y de ofrecer resultados comparables a los de un avanzado sistema basado en reglas. Es cierto, no obstante, que precisan de un corpus de entrenamiento razonablemente grande, aunque es precisamente aquí donde reside su flexibilidad, ya que para adaptarlo a un nuevo idioma solamente habrá que redefinir las categorías y proporcionarle un nuevo corpus de entrenamiento.

Por supuesto, también existen métodos híbridos, que intentan combinar las ventajas de las reglas con las de los métodos estocásticos. En (Màrquez, 1999) se recoge una descripción general más completa.

En este artículo presentamos un método híbrido de análisis morfosintáctico basado en la combinación de un conjunto reducido de reglas lingüísticas y un modelo estadístico que emplea pentagramas, es decir, secuencias de cinco categorías gramaticales consecutivas. Este analizador es capaz de determinar no sólo la función gramatical de cada palabra sino también información adicional como género y numero, cuando corresponda, o tiempo, modo, número y persona de las formas verbales. Además, explicaremos la metodología empleada para crear un nuevo corpus de texto en gallego correctamente etiquetado en el nivel morfosintáctico. Con este objeto, en el apartado 2 se realiza una descripción de las características del texto seleccionado; en la sección 3 se enumera el conjunto de categorías léxicas empleadas; en la sección 4 se describe el proceso iterativo seguido para la creación del corpus de texto analizado con el que se entrenará el modelo estadístico de lenguaje; en el apartado 5 se exponen los dos modelos estadísticos extraídos del corpus, el primero referido a las probabilidades contextuales, es decir, a la probabilidad de aparición de una serie de categorías consecutivas, y el segundo referido a las probabilidades léxicas, es decir, la probabilidad de que una determinada palabra cumpla cierta función en la frase; finalmente en la sección 6 se presenta el procedimiento general seguido para el análisis morfosintáctico, discutiéndose diferentes algoritmos para la aplicación de los modelos anteriormente mencionados. Por supuesto, también se presentan los resultados obtenidos (apartado 7) y una serie de conclusiones y futuras líneas de trabajo (sección 8).

2. Descripción del corpus textual

Los métodos estadísticos precisan una cantidad de texto relativamente grande para poder obtener unas estimaciones aceptables de las probabilidades de ocurrencia de las secuencias de categorías. Mientras que para otros idiomas se encuentra algún corpus disponible (por ejemplo, el "Penn Treebank" y el "Corpus Brown" para el inglés, el corpus "NEGRA" para el alemán o el corpus "LexEsp" en castellano), en gallego no existen todavía recursos disponibles de este tipo. Por este motivo ha sido necesario crear un corpus de texto en gallego analizado morfosintácticamente que pueda ser empleado para entrenar las probabilidades del modelo del lenguaje o incluso servir como referencia para otros trabajos de investigación.

Para su elaboración decidimos emplear texto periodístico no restringido. Se trata de un texto complejo con estructuras intrincadas, que engloba artículos de muy diversos autores y diferente temática, con una gran variedad de estilos. En el futuro pensamos incluir otros tipos de texto.

3. Etiquetas morfosintácticas

A continuación se describe el conjunto de categorías morfosintácticas que hemos utilizado en el etiquetado del corpus, que son las mismas que emplea internamente nuestro sistema de conversión texto-voz.

En general, consideramos género y número en aquellas categorías en que tiene sentido. Los posibles valores son *masculino*, *femenino*, *neutro* para el género y *singular* o *plural* para el número. También se utiliza la etiqueta *ambiguo* en algunos casos, por ejemplo con ciertos pronombres como los reflexivos y también en algunos nombres propios.

Las categorías utilizadas son:

- Artículo determinado e indeterminado, y sus correspondientes contracciones con preposiciones, conjunciones o indefinidos.
- Determinantes y pronombres demostrativos, con sus contracciones con preposición, indefinido o preposición+indefinido.

- Pronombres y determinantes posesivos, junto con el posesivo distributivo.
- Numerales (determinantes y pronombres) cardinales, ordinales, partitivos, multiplicativos y colectivos.
- Pronombres personales tónicos y átonos (reflexivos, acusativos, dativos, etc.). También diferentes secuencias de pronombres.
- Adverbios y locuciones adverbiales de lugar, tiempo, cantidad, modo,afirmación, negación, de duda,...
- Preposiciones y locuciones prepositivas de lugar, tiempo, cantidad, modo, causa y condicional.
- Pronombres relativos, exclamativos e interrogativos.
- Conjunciones coordinadas copulativas, disyuntivas y adversativas. También locuciones conjuntivas copulativas.
- Conjunciones y locuciones conjuntivas subordinantes de diversos tipos: causal, concesiva, consecutiva, condicional, comparativa, locativa, temporal, modal, etc.
- Nombre, nombre propio y adjetivo.
- Verbo y perífrasis verbal.

En el caso de los verbos, se presenta un análisis detallado que incluye información de persona, número, tiempo, modo, conjugación, infinitivo del verbo, segunda forma del artículo o pronombres enclíticos cuando proceda, presentando además en estos casos la "reconstrucción" de la forma verbal sin el pronombre o artículo.

En total suman unas 300 etiquetas diferentes (500 si se considera el análisis verbal completo).

4. Procedimiento de creación del corpus

Dado que un etiquetado totalmente manual del corpus sería una labor larga, ardua y tediosa, este proceso se ha realizado por medio de una técnica iterativa de *bootstrapping*. En primer lugar se etiqueta automáticamente una pequeña cantidad de texto empleando únicamente un conjunto reducido de reglas lingüísticas. Tras su posterior revisión manual por parte de dos lingüistas competentes, empleando herramientas que previenen la introducción de ciertos errores, se construye un modelo estadístico inicial, el cual se incorporará al análisis de los sucesivos textos. De esta forma, el procedimiento consiste en una serie de etapas de etiquetado automático con la última versión del modelo estadístico, revisión manual y actualización del modelo con el texto correctamente etiquetado. Además, también se emplean las diferencias entre el texto analizado manualmente y el generado automáticamente como un medio para corregir errores que se hayan pasado por alto en la revisión manual y para inferir reglas que corrijan aquellos errores sistemáticos que cometa el conversor. Esta técnica tiene la ventaja de que la revisión manual es cada vez más sencilla, al estar más libre de errores el texto etiquetado automaticamente. De hecho, cuando se introdujo el primer modelo estadístico no se advirtió de ello a los lingüistas que hacían la revisión y, sin embargo, en seguida se percataron de que la fiabilidad del etiquetado automático se había incrementado notablemente.

En estos momentos, nuestro corpus consta de unas 360.000 palabras etiquetadas según esta metodología. Nuestra intención es alcanzar el millón de palabras a finales de este año.

5. Modelos estadísticos

Nuestra metodología se basa en la obtención de dos modelos probabilísticos a partir del corpus: el modelo contextual, que contiene las probabilidades de aparición de las diferentes secuencias de categorías, y el modelo léxico, que proporciona las probabilidades de que cada palabra se etiquete con una categoría en concreto.

Para estimar el modelo contextual utilizamos las herramientas del CMU–Cambridge Toolkit (Clarkson y Rosenfeld, 1997), que permiten estimar las probabilidades de ocurrencia de n–gramas. En nuestro caso, obtenemos un modelo basado en grupos de cinco etiquetas (pentagramas), tomando como entrada una secuencia de categorías, obtenida a partir de una porción del corpus de unas 260.000 palabras, incluyendo signos de puntuación, que llamamos corpus de entrenamiento.

Es necesario suavizar estas probabilidades para permitir la posibilidad de encontrar casos diferentes a los hallados en el corpus de entrenamiento. Con este objeto utilizamos la técnica de *back–off* y descuento Witten–Bell,

habiéndose realizado experimentos con otras estrategias, como descuento absoluto, lineal o Good–Turing[1], obteniendo peores resultados, tanto a nivel de perplejidad[2] del modelo como en tasa de aciertos en el etiquetado.

Con el tamaño actual del corpus de entrenamiento, la utilización directa de un conjunto de categorías tan amplio como el descrito anteriormente hubiera llevado con frecuencia a una estimación incorrecta, poco fiable, de las probabilidades de los n-gramas, a causa de la imposibilidad de cubrir todas las posibles combinaciones de categorías con un texto de ese tamaño. Por tanto, y de manera similar a la efectuada en otros trabajos, como por ejemplo (Tzoukermann y Radev, 1997)(Ezeiza et al., 1998)(Oliver, 1998), se ha realizado una simplificación del conjunto de categorías, agrupando aquellas categorías afines, o cuya distinción pensamos que no aporta información relevante para realizar la desambiguación morfosintáctica. Además, este proceso de reducción es totalmente reversible.

Cuadro 1: Conjunto de categorías simplificado

ADV	Adverbio
ADX	Adjetivo
APERTURA_EXCLA	Signo de apertura de exclamación
APERTURA_INTERR	Signo de apertura de interrogación
COM	Coma
CONX_COORD	Conjunción coordinada
CONX_SUBORD	Conjunción subordinada
DET	Determinante
DET_NUME	Determinante numeral
DET_POSE	Determinante posesivo
EXCLAM	Pronombre exclamativo
INTERR	Pronombre interrogativo
INTERXE	Interjección
NOME	Nombre
PECHE_EXCLA	Signo de cierre de exclamación
PECHE_INTERR	Signo de cierre de interrogación
PREPO	Preposición
PRON	Pronombre
PRON_NUME	Pronombre numeral
PRON_POSE	Pronombre posesivo
PRON_PROC	Pronombre proclítico
PUNT	Puntuación de final de frase
RELAT	Pronombre relativo
VERBO	Verbo
XERUND	Gerundio

De forma simplificada, el conjunto de categorías utilizado finalmente para la estimación de los modelos es el que se presenta en el cuadro 1. En realidad, cada categoría de esta tabla que admite variaciones de género y número es desdoblada en tantas etiquetas como sean necesarias para tener en cuenta dichas variaciones y, de esta forma, considerar la concordancia de género y número entre palabras sucesivas. Para llegar al cuadro 1 se han realizado una serie de simplificaciones que describimos a continuación:

- Agrupar artículo determinado, indeterminado y determinantes demostrativos e indefinidos en una única categoría, determinante genérico.
- Los pronombres se reducen a dos tipos, uno genérico y otro proclítico, que aparecerá siempre antes de un verbo.
- Las categorías determinante / pronombre posesivo y numeral se mantienen.
- Se agrupan nombres comunes y propios.
- Se considera un único tipo de adverbio genérico.
- Las locuciones se tranforman en su categoría "madre": esto es, las prepositivas en preposición, las conjuntivas en conjunción, etc.
- Las conjunciones y locuciones conjuntivas se dividen en dos tipos, coordinante y subordinante.
- A efectos de aprovechar la información de concordancia, se utiliza género y número en determinantes, nombres, pronombres y adjetivos, así como el número en las formas verbales.

Como resultado de todo este proceso las 300 etiquetas iniciales se reducen a 53.

En lo que respecta al modelo léxico, para su construcción utilizamos el concepto de clase de ambigüedad (Cutting et al., 1992) (Tzoukermann y Radev, 1997), conjuntos de palabras del corpus de entrenamiento agrupadas según todas las posibles etiquetas que se le pueden asignar a priori como, por ejemplo, *pronombre–determinante–preposición*, *relativo–subordinante*, etc. De este modo, el número de ocurrencias de cada clase de ambigüedad es suficiente para estimar las probabilidades de forma aproximada, sin necesidad de estimar directamente la probabilidad conjunta de cada categoría y palabra en concreto.

[1]Una revisión de esta problemática y de los diferentes métodos de suavizado puede encontrarse en (Nivre, 2000)

[2]Medida de calidad de los modelos de lenguaje, tanto del modelado de los datos de entrenamiento como su capacidad de generalización ante un texto no presente en el entrenamiento

El modelo léxico se ha obtenido a partir del mismo corpus de entrenamiento que el modelo contextual, agrupando las diferentes clases de ambigüedad y estimando las probabilidades a partir de las frecuencias relativas de aparición de las diferentes etiquetas dentro de cada clase. Para no descartar posibilidades que no aparecen recogidas en nuestro corpus, se han asignado probabilidades muy bajas a aquellas categorías dentro de la clase de las que no se ha contado ninguna ocurrencia en el texto de entrenamiento. Para tratar de paliar este problema, en la construcción de las clases de ambigüedad también se ha utilizado el conjunto de categorías reducido, esta vez sin tener en cuenta variación de género ni número, ya que, al considerar las palabras de forma aislada, pensamos que esta información no es relevante. En total, se obtienen 192 clases de ambigüedad diferentes, que contienen desde un mínimo de dos hasta un máximo de ocho categorías distintas. Si se tuvieran en cuenta género y número, este número crecería hasta 20.

Ambos modelos, contextual y léxico, se almacenan en memoria mediante estructuras de tipo *hash*, lo que durante el proceso de etiquetado automático permite una obtención de las probabilidades del modelo de forma eficiente desde el punto de vista computacional.

6. El algoritmo de etiquetado

Podemos dividirlo en dos etapas: una primera de anotación, en la que se asigna a cada palabra todas sus posibles etiquetas y una segunda fase de desambiguación, que consiste en escoger la etiqueta más probable entre todas las posibilidades.

6.1. Anotación

Para encontrar todas las posibles categorías a las que a priori puede pertenecer cada palabra disponemos de una serie de diccionarios o tablas obtenidos de forma independiente al corpus de entrenamiento. En concreto utilizamos:

- Un diccionario básico, que contiene palabras pertenecientes a categorías léxicas cerradas: preposiciones, determinantes, pronombres, indefinidos, conjunciones, adverbios, contracciones, así como los numerales e interjecciones más comunes. En total este diccionario consta de aproximadamente 900 palabras.
- Diccionarios de adjetivos y nombres (30.000 y 150.000 respectivamente), teniendo en cuenta variaciones de género y número. No se incluyen nombres propios.
- Una tabla con las locuciones (preposicionales, adverbiales y conjuntivas) más comunes (aproximadamente 600).
- Tablas de abreviaturas y acrónimos más utilizados.
- Los verbos se detectan empleando una descomposición en pares raíz–desinencia, asignando a cada raíz un modelo de conjugación, que conlleva la utilización de un determinado conjunto de desinencias (existen 92 modelos diferentes). Por lo tanto, empleamos dos diccionarios, uno de raíces, que consta de unas 8500 entradas, y otro de terminaciones o desinencias (684). En total, nos permite analizar más de 350.000 formas verbales diferentes, sin necesidad de utilizar diccionarios tan grandes. Además, también se detectan los pronombres enclíticos, bien de forma aislada o formando conglomerados de dos o más pronombres. Para más información sobre el análisis verbal, puede consultarse (González et al., 2002).
- Tabla que contiene la descripción de las estructuras consideradas como perífrasis verbales: *verbo + infinitivo*,*verbo + participio*, *verbo + gerundio*, *verbo + preposición + infinitivo*, etc...

En la identificación de los nombres propios sirve de gran ayuda el conocimiento de si la palabra empieza o no por mayúscula. Además, se estudia la terminación de las palabras para detectar algunos casos característicos (por ejemplo, los adverbios de modo terminados en *mente*)

6.2. Desambiguación

Para decidir cuál es la categoría correcta para cada palabra, entre todas las posibilidades de la clase de ambigüedad, utilizamos un esquema híbrido similar al descrito en (Ezeiza et al., 1998) para el euskera. Así, los modelos estadísticos, se combinan con un conjunto reducido de reglas lingüísticas (unas cincuenta) que, teniendo en cuenta el contexto y la propia naturaleza de cada palabra, intentan reducir la ambigüedad descartando aquellas

categorías que se consideran improcedentes en determinadas situaciones.

Posteriormente, se realiza la conversión de las categorías morfosintácticas al conjunto simplificado que se utiliza en la obtención de los modelos probabilísticos. Con estos modelos se estima la secuencia de categorías más probable para cada frase. Para determinar dicha secuencia se han probado tres algoritmos distintos que discutiremos posteriormente.

Además de la transformación de categorías es necesario realizar otra serie de acciones, como el desdoblamiento de las contracciones en *preposición* o *conjunción* más *determinante* o *pronombre* según proceda, la agrupación o asimilación a una única categoría en el caso de las perífrasis verbales (*verbo*) y locuciones (*preposición*, *conjunción* o *adverbio*, dependiendo del tipo de locución) y la separación de la segunda forma del artículo de las formas verbales. En el caso de los participios y adjetivos, se admite la posibilidad de que puedan aparecer desempeñando funciones propias de sustantivo, de ahí que también se les asigne a la clase de ambigüedad *adjetivo-adjetivo sustantivado*. Algo parecido sucede con los infinitivos, considerando en este caso la clase de ambigüedad *verbo–verbo sustantivado*, diferenciándola de la clase *verbo–nombre*.

En adelante denotaremos como $\mathbf{C} \equiv \{c_0, \ldots, c_N\}$ el conjunto de todas las posibles categorías consideradas en el modelo estadístico, y como $\mathbf{F} \equiv \{p_0, \ldots, p_L\}$ una frase de longitud L palabras. El objetivo es encontrar la secuencia correcta de etiquetas $\widehat{C} = \{c_0, \ldots, c_L\}$ para la frase de entrada $\mathbf{F}$.

Se define la clase de ambigüedad a_i para cada palabra p_i como

$$a_i = \{c_0, \ldots, c_{N_i}\} \quad c_i \in \mathbf{C} \quad 1 \leq N_i \leq N$$

y la probabilidad contextual $P(c_{ji}/c_{i_{-1}}, \ldots, c_{i_{-4}})$ como la probabilidad de que la palabra p_i tenga la categoría c_j suponiendo que las 4 palabras anteriores pertenecen a las categorías $c_{i_{-1}}, \ldots, c_{i_{-4}}$ respectivamente.

Definimos las probabilidades léxicas para cada palabra $P(c_{ji}/a_i)$ como la probabilidad de que la palabra p_i tenga la categoría c_j perteneciente a su clase de ambigüedad a_i.

El caso de palabras desconocidas o fuera de vocabulario se trata como se describe a continuación:

- Las palabras que, en el paso anterior de anotación, no se han encontrado en ningún diccionario ni se han identificado como verbo o numeral, se asignan a la clase de ambigüedad *nombre–adjetivo*, teniendo en cuenta todas las posibles combinaciones de género y número.
- Para las probabilidades contextuales, en el caso de pentagramas no vistos en el texto de entrenamiento, su probabilidad se obtiene recursivamente a partir de la del tetragrama (o, en caso de no existir éste, del trigrama, bigrama o unigrama sucesivamente) y su *back–off* correspondiente.
- Para las clases de ambigüedad no vistas en el entrenamiento, las probabilidades de cada categoría de la clase se aproxima por la del unigrama correspondiente.

Finalmente describimos los algoritmos empleados para detectar la secuencia de categorías más probable.

Método A (Algoritmo de Viterbi)

Método clásico de programación dinámica que obtiene la secuencia de estados de mayor probabilidad. En nuestro problema, las probabilidades de transición entre estados están representadas en el término $P(c_{ji}/c_{i_{-1}}, \ldots, c_{i_{-4}})$ de la ecuación 1 (probabilidad contextual) , mientras que el término $P(c_{ji}/a_i)$ representa la probabilidad que la palabra actual tenga una determinada categoría entre las presentes en su clase de ambigüedad. En nuestro caso, utilizando un modelo contextual basado en pentagramas, el número de estados es el número de posibles combinaciones o secuencias de cuatro categorías.

$$\widehat{C} = \arg\max_{\substack{c_{ji} \in a_i \\ j=0\ldots N_i}} \prod_{i=1\ldots L} P(c_{ji}/c_{i_{-1}}, \ldots, c_{i_{-4}}) P(c_{ji}/a_i) \tag{1}$$

Método B *Ventana deslizante*: Se desplaza a lo largo de la frase una ventana de observación de cinco palabras. Se decide la categoría correcta en la palabra central maximizando el producto de la probabilidad contextual (teniendo en cuenta todos los posibles pentagramas construidos a partir de las clases de ambigüedad de las cinco palabras de la ventana, primer elemento en la ecuación 2) por la probabilidad léxica de cada categoría

en la clase de ambigüedad de la palabra central (segundo término en la ec. 2).

$$\widehat{c_i} = \arg\max_{\substack{c_{ji}\in a_i \\ j=0...N_i}} P(c_{ji-2},\ldots,c_{ji+2})P(c_{ji}/a_i) \quad (2)$$

Este método va decidiendo la categoría más probable palabra a palabra. Sin embargo, para maximizar la probabilidad contextual, utiliza todas las posibles categorías de las palabras anteriores, no sólo la decidida previamente como correcta. Esta consideración permite al método recuperarse de posibles errores anteriores.

Método C: Refinamiento del método anterior, pero considerando las probabilidades léxicas de todas las palabras de la ventana (ecuación 3).

$$\widehat{c_i} = \arg\max_{\substack{c_{ji}\in a_i \\ j=0...N_i}} P(c_{ji-2},\cdots,c_{ji+2}) \prod_{k=-2...2} P(c_{ji+k}/a_{i+k}) \quad (3)$$

Una vez obtenida la secuencia de categorías más probable para la frase de entrada utilizando uno de los 3 métodos anteriores, tiene lugar la transformación inversa del conjunto de categorías reducido al conjunto original. Es preciso mencionar que en algunos casos se pueden generar ambigüedades. Por ejemplo, en ocasiones, en palabras que pueden funcionar como diferentes tipos de adverbio, no se determina unívocamente de qué tipo concreto son.

7. *Evaluación del algoritmo: Resultados experimentales*

Como se ha mencionado anteriormente, el desarrollo y depuración de nuestro analizador estadístico se ha desarrollado de forma paralela al proceso de corrección manual y creación del corpus etiquetado. Los primeros resultados de los modelos estadísticos han comenzado a ser razonables a partir de un texto de entrenamiento de unas 50.000 palabras. En la figura 1 puede verse la evolución de la precisión en el etiquetado en función de la cantidad de texto de entrenamiento. Como es habitual en este tipo de sistemas, llega un momento en que para lograr pequeñas mejoras se requiere una gran cantidad de texto analizado adicional. Los resultados aquí presentados se han obtenido al etiquetar un texto de unas 25.000 palabras, no incluido en el material de entrenamiento.

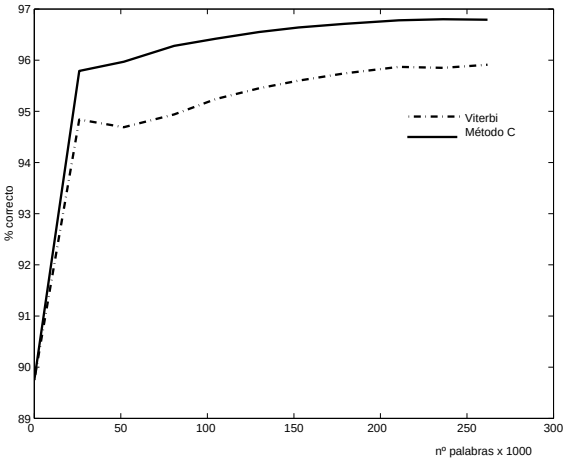

Figura 1: Precisión del etiquetado vs. tamaño del corpus de entrenamiento.

A continuación, en el cuadro 2 detallamos la tasa de acierto para los distintos métodos propuestos, teniendo en cuenta por separado categoría, género y número. Los resultados que se muestran han sido obtenidos combinando los modelos estadísticos con las reglas lingüísticas y también utilizando los dos métodos por separado (para observar la mejora introducida por la combinación de los dos). En caso de utilizar únicamente las reglas, la categoría final se decide de un modo aleatorio entre las presentes en la clase de ambigüedad. Se puede observar como la utilización única de reglas proporciona peores tasas de acierto que los modelos estadísticos y éstos peores resultados que la combinación de reglas y modelos estadísticos. También se comprueba que el método C proporciona mejores prestaciones que A y B.

Cuadro 2: Resultados

Reglas	Método	Categoría	Género	Número
sí	-	89.75	85.84	88.49
no	A	93.77	95.96	97.63
	B	95.64	97.44	98.42
	C	95.89	97.45	98.43
sí	A	95.91	96.30	97.80
	B	95.54	97.60	98.59
	C	**96.79**	**97.63**	**98.59**

8. Conclusiones y trabajo futuro

En este artículo hemos presentado la metodología seguida para la elaboración de un corpus de texto en gallego analizado morfosintácticamente, por medio de un procedimiento iterativo en el que se intercalan fases de análisis automático con revisiones manuales. Actualmente, el corpus consta de casi 400.000 palabras de texto periodístico no restringido, estando previsto llegar hasta el millón a finales de año.

Del texto analizado se han obtenido dos modelos probabilísticos: uno, referido a la probabilidad de ocurrencia de cada secuencia de categorías, y otro referido a la probabilidad de que cada palabra tenga cierta categoría en el contexto de la frase. Posteriormente, se han desarrollado diferentes algoritmos para la detección de la secuencia de categorías más probable, obteniéndose unos resultados que, aunque no son directamente comparables a los reflejados en otros trabajos sobre lenguas diferentes, son realmente prometedores.

Como líneas futuras, exploraremos las posibilidades de un esquema para completar y depurar las reglas lingüísticas automáticamente, de manera similar a las reglas de transformación en (Brill, 1995). Más directamente relacionado con la aplicación a las tecnologías del habla, mejoraremos el modelado prosódico de nuestro sistema de conversión texto-voz y aplicaremos el análisis morfosintáctico estadístico a un sistema de reconocimiento de habla continua.

Bibliografía

Brants, T. 2000. TnT – a statistical part-of-speech tagger. En *Proceedings of the 6th Applied NLP Conference, ANLP-2000*, Seattle, WA.

Brill, E. 1995. Transformation–based Error–driven Learning and Natural Language Processing: A Case Study in Part–of–speech Tagging. *Computational Linguistics*, 21(4):543–565.

Clarkson, Philip y Ronald Rosenfeld. 1997. Statistical language modeling using the CMU-cambridge toolkit. En *Proc. Eurospeech '97*, páginas 2707–2710, Rhodes, Greece.

Cutting, D., J. Kupiec, J. Pederson, y P. Sibun. 1992. A Practical Part–of–speech Tagger. En *Proceedings of the 3rd Conference on Applied Natural Language Processing, ANLP*, páginas 133–140. ACL.

Daelemans, W., J. Zavrel, P. Berck, y S. Gillis. 1996. MBT: A Memory–Based Part–of–speech Tagger Generator. En *Proceedings of the 4th Workshop on Very Large Corpora*, páginas 14–27, Copenhagen, Denmark.

Ezeiza, N., I. Alegria, J. M. Arriola, R. Urizar, y I. Aduriz. 1998. Combining stochastic and rule-based methods for disambiguation in agglutinative languages. En Christian Boitet y Pete Whitelock, editores, *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, páginas 379–384, San Francisco, California. Morgan Kaufmann Publishers.

González, Manuel, Carmen García Mateo, Eduardo Rodríguez Banga, y Elisa Fernández Rei. 2002. *Diccionario de verbos galegos LAVERCA*. Edicións Xerais de Galicia, España.

Màrquez, L. 1999. *Part–of–Speech Tagging: A Machine–Learning Approach based on Decision Trees*. Phd. Thesis, Dep. Llenguatges i Sistemes Informàtics. Universitat Politècnica de Catalunya.

Nivre, Joakim. 2000. Sparse Data and Smoothing in Statistical Part-of-Speech Tagging. *Journal of Quantitative Linguistics 7(1)*, páginas 1–17.

Oliver, Dan Tufis. 1998. Tagging romanian texts: a case study for qtag, a language independent probabilistic tagger.

Ratnaparkhi, Adwait. 1996. A maximum entropy model for part-of-speech tagging. En Eric Brill y Kenneth Church, editores, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Somerset, New Jersey, páginas 133–142.

Tzoukermann, Evelyne y Dragomir R. Radev. 1997. Tagging french without lexical probabilities - combining linguistic knowledge and statistical learning.