OBTENCIÓN AUTOMÁTICA DE MARCOS DE SUBCATEGORIZACIÓN VERBAL A PARTIR DE TEXTO ETIQUETADO: EL SISTEMA *SOAMAS*

Juan Monedero, José C. González, José M. Goñi,
Carlos A. Iglesias, Amalio F. Nieto
E.T.S.I. Telecomunicación
Universidad Politécnica de Madrid

Departamento de Ingeniería de Sistemas Telemáticos
E.T.S.I.Telecomunicación
Universidad Politécnica de Madrid
Ciudad Universitaria s/n
28040 Madrid
tel: 5495700
fax: 5432077
mail: jgonzalez@dit.upm.es

Tema: Adquisición automática de conocimiento lingüístico.

Palabras clave: adquisición de conocimiento lingüístico, recursos léxicos, técnicas basadas en corpus, análisis sintáctico.

Este trabajo presenta una herramienta para la obtención automática de marcos de subcategorización de los verbos españoles a partir de textos etiquetados morfosintácticamente. Dada una oración, el sistema desarrollado trata de identificar y analizar el sintagma verbal a fin de reconocer los complementos presentes (junto con la proposición asociada si es el caso), formas impersonales, reflexividad, auxiliariedad, etc.

1 Introducción

1.1 Los marcos de subcategorización

Marcos de subcategorización son las diferentes estructuras sintácticas que un verbo admite: qué tipo de complementos acepta, formas impersonales, formas reflexivas, combinación con preposiciones, etc.

Tomemos como ejemplo las dos siguientes oraciones :

```
      Juan come patatas.
(*)   Juan corre patatas.
```

La segunda oración no es correcta en español. El verbo `correr`, intransitivo, no acepta un complemento directo. En otras palabras, el complemento directo no es un *marco de subcategorización* del verbo `correr`. Haciendo una osada analogía computacional, los marcos de sub-

categorización de un verbo son como los argumentos de una función. La función `comer` admite como argumento un complemento directo, encarnado en este caso por el sustantivo `patatas`. Sin embargo, `correr` no admite dicho argumento. Como consecuencia, la frase no es sintácticamente correcta.

Frecuentemente, y este caso no es la excepción, la frontera entre sintaxis y semántica es estrecha y difusa. Tomemos ahora como ejemplo estas otras dos oraciones:

a) `Juan piensa en tí.`

b) `Juan lee en casa.`

El verbo `pensar` admite el marco *complemento preposicional*, con la preposición `en`. Sin embargo, `leer` no subcategoriza para dicha preposición. En cierto modo, la preposición es una *extensión* del verbo `pensar` ([Hallebeek 92; p126]). Entre ambos se produce una sinergia, cuyo resultado es un significado mayor que la simple concatenación, tal y como ocurre en la segunda oración: `en` proporciona información acerca de dónde se realiza la acción, pero no se encuentra ligado al verbo de una forma tan estrecha como en la primera oración.

Veamos un ejemplo más :

c) `Juan piensa en tí en casa.`

Se aprecia ahora mucho más claramente la diferente función que realizan los dos `en`. Llamaremos al primero de los casos *complemento preposicional*. Es un marco de subcategorización, que además se parametriza por la preposición correspondiente (en este caso `en`). El segundo es un *complemento adverbial*, y no tiene la categoría de marco. Los conceptos clave son *argumento* y *adjunto*[1]. Los marcos de subcategorización describen los *argumentos* que un verbo puede tomar. Los complementos adverbiales son simplemente *adjuntos*. Un complemento adverbial puede aplicarse a cualquier verbo, ya que sólo modifica las circunstancias en que se desarrolla la acción. Y en toda acción puede describirse un entorno de tiempo, lugar, etc. Es diferente que cierto tipo de acciones *requieran*, por ejemplo, producirse *directamente sobre algo o alguien* (como sería el caso de los verbos transitivos).

1.2 Por qué nos interesan y por qué obtenerlos de forma automática.

La utilidad de una herramienta de las características de *SOAMAS* tiene dos vertientes. En primer lugar, como forma de obtener conocimiento lingüístico que sirva de base para sistemas de lenguaje natural basados en conocimiento. En este sentido, la información de subcategorización puede ser utilizada por los analizadores sintácticos para reducir el número de posibles análisis. Tradicionalmente esta información se recoge y codifica en forma no automática, lo cual induce problemas como su volumen escaso y reducida cobertura[2], que además ha costado un ingente trabajo recopilar ([Manning 93], [Brent 91], [Basili *et al.* 93])[3]. Un ejemplo es la subcategorización usada por Hallebeek en

1. [Manning 93].

[Hallebeek 92]. *SOAMAS* puede ser útil usado también en conjunción con el estudio manual, ya que ambos métodos se complementan: puede proporcionarse un gran volumen de información que complete una base ya preparada, de forma que se consiga eliminar errores arbitrarios y por tanto una mayor relevancia.

Desde el punto de vista semántico, el conocimiento de los marcos válidos para un verbo es esencial para atribuir funciones semánticas a cada uno de sus complementos.

Todavía en este sentido, *SOAMAS* es útil en sistemas de generación automática, que utilizan subcategorización para decidir entre las posibles opciones para una misma o similar semántica ([Manning 93]). E.g.

```
        Pienso en tí.
(*)     Pienso a tí.
```

La subcategorización de `pensar` sugeriría utilizar la preposición `en` como opción mucho mejor que `a`.

Un segundo aspecto hace referencia al conocimiento como fin y no como medio (en oposición al caso anterior):

- Los diccionarios suelen carecer de información de subcategorización[1]. Además, no suelen recoger usos especializados del lenguaje (jergas profesionales, por ejemplo) que *SOAMAS* podría recoger si se le proporcionaran textos adecuados : manuales, revistas técnicas, etc.
- La obtención automática permite detectar usos que pasan frecuente o arbitrariamente inadvertidos.
- El mantenimiento y la actualización de bases léxicas se simplifica notablemente usando herramientas automáticas.
- Finalmente, hay aspectos que simplemente todavía no han sido estudiados, como la predominancia de uno u otro orden relativo de los complementos entre los posibles sintácticamente correctos ([Hallebeek 92; p.115]).

1.3 El estado del arte

El presente trabajo se inspira en los realizados por Brent ([Brent 91], [Brent 93]) y Manning ([Manning 93]), no teniendo constancia de que se hayan realizado trabajos en esta línea para el lenguaje español.

---

2. [Church *et al.* 90] ponen de manifiesto la selectividad y distorsión que produce la recopilación humana, como, por ejemplo, la anotación de palabras y usos infrecuentes y la no inclusión de otros muy comunes.

3. Un ejemplo es la subcategorización verbal recopilada por Hallebeek en [Hallebeek 92; p. 280-293]].

1. Hay algunas excepciones, tales como el *Oxford Advanced Learner's Dictionary,* que ya en su edición de 1948 recogía *patrones verbales* para la correcta construcción sintáctica en inglés.

[Brent 91] describe un programa cuya entrada es un corpus sin ningún tipo de marca. Del análisis de este corpus obtiene un lista de verbos subcategorizados para cinco marcos diferentes. El pilar central del trabajo de Brent es obtener una fiabilidad muy alta en la obtención de marcos. Busca obtener información sin apenas errores, a costa, sin embargo, de tener una eficiencia muy baja en el uso del corpus: argumenta que conviene esperar a aquellos casos que no ofrezcan dudas ni ambigüedades[1]. Efectivamente, los errores que se derivan del método de trabajo de Brent se encuentran entre un 0,5% y un 3%, según el propio Brent.

Dos años más tarde ([Brent 93]), es capaz de reconocer seis marcos, tan sólo uno más que en el anterior. Esto pone de manifiesto el principal problema al que se enfrenta: hay muchos marcos para los que simplemente no hay *pistas* fiables ([Manning 93]). Por tanto, Brent se ve obligado a ceñirse a unos pocos marcos para los que sí existen dichas pistas.

Otro problema del enfoque de Brent, si bien más secundario, es el de la bajísima eficiencia en el uso del corpus.Tanto la identificación de subcategorización como la de verbos principales desprecian un altísimo porcentaje de la información potencialmente disponible.

En [Manning 93] se propone un punto de vista alternativo. Su argumento es que la alta fiabilidad es un callejón sin salida que limita el tipo de marcos que pueden estudiarse. Busquemos, por tanto, una forma de trabajo diferente. En concreto, sugiere obtener información razonablemente *ruidosa*, pero más completa. En otras palabras, vamos a usar métodos de detección mucho menos fiables que los de Brent, pero que puedan usarse para muchos tipos de marcos, evitando ceñirnos a los escasos seis propuestos por Brent. Ahora bien, ¿cómo reducir el ruido que acompaña a esta información? [Brent 93] describe un método estadístico que es adoptado también por Manning. Brevemente, el método se basa en que los errores pueden ser eliminados observando qué marcos aparecen con un verbo en una frecuencia razonablemente superior a lo que pudiera considerarse casualidad (complementación adverbial) o errores en la detección.

Con este filtrado estadístico posterior podemos permitirnos utilizar un etiquetador estocástico automático que nos marque el texto, asumiendo que su carácter probabilístico produce errores. Para etiquetar un texto existen dos opciones. Una es etiquetarlo en forma no automática o semi-automática. El etiquetado semi-automático requiere que una persona opte entre las varias opciones que son sugeridas por un programa para aquellas palabras que son sintácticamente ambiguas. El tiempo que se requeriría para un corpus grande sería del mismo orden que el necesario para un hipotético etiquetado exclusivamente con información de subcategorización verbal. Además, preparar un corpus de varios millones de palabras en forma semi-automática no parece la opción más sensata.

Es más lógico inclinarse a favor de un etiquetador automático. Actualmente nuestro grupo trabaja en la adaptación al español de un *etiquetador estocástico*[2]. Un etiquetador estocás-

1. Se basa principalmente en la flexión de caso de los pronombres en inglés y un método muy fiable de detección de verbos en texto no marcado: el *Filtro de Caso* ([Rouvret *et al.* 80]).

tico es, como su propio nombre indica, un sistema capaz de concatenar información a las palabras de un corpus (las *etiquetas*), de acuerdo con un funcionamiento probabilístico. Se trata de un sistema ciertamente complejo. Brevemente: basándose en la etiqueta que se ha dado a una o dos palabras anteriores, se analiza la siguiente; dentro de las posibles interpretaciones que se pueda dar a una *palabra gráfica* (es decir, al conjunto de caracteres), se opta por etiquetar esta última según las diferentes *probabilidades de transición* entre las etiquetas anteriores, como se ha dicho, y las citadas diferentes posibilidades existentes. Para obtener estas diferentes probabilidades de transición, es necesario someter al etiquetador a un *entrenamiento*, enseñándole las opciones correctas mientras lee un conjunto de textos. La teoría subyacente está basada en *cadenas de Markov.*

2 Criterios de diseño de *SOAMAS*

2.1 El etiquetador estocástico y el procesado estadístico.

Como ya se ha dicho, *SOAMAS* está inspirado en los trabajos de Brent y Manning. En particular, se ha preferido el enfoque de Manning. *SOAMAS*, en una balanza de compromiso entre exactitud y volumen de información obtenida, se inclina por el segundo concepto, llegando algo más lejos que Manning. Consecuencia de este enfoque, como ya se ha comentado, es la posibilidad de utilización de un etiquetador automático y la necesidad del post-procesado estadístico.

SOAMAS precisa que el texto de entrada sea texto etiquetado. Idealmente, este texto debe ser constitutivo de un *corpus*, a ser posible lo suficientemente extenso, variado y equilibrado como para ser considerado representativo del español. Por el momento se dispone tan sólo de unas 10.000 palabras etiquetadas por Aurora Martín de Santa Olalla como parte de su tesis doctoral [Martín 94]. Por otra parte, tal como se mencionó previamente, se está trabajando en la adaptación de un etiquetador estocástico. En concreto, se trata de un software de libre distribución realizado por Julian Kupiec, [Kupiec 92] y [Cutting *et al.* 92]. En estos artículos, Kupiec argumenta que su sistema puede ser trivialmente modificado para otros lenguajes de los que se disponga de una base léxica. Por desgracia, se ha visto que se trataba de una presunción exageradamente optimista. En nuestro grupo se dispone de una base léxica aceptable, que se está mejorando continuamente. Los trabajos de adaptación del etiquetador sobrepasan ya los seis meses, pero los resultados son alentadores. En concreto, se cree que el nivel de etiquetado que puede llegar a proporcionarse es superior al necesario en *SOAMAS* : categorías gramaticales, auxiliariedad e incorporación pronominal verbal, y casos pronominales. Como digo, los resultados obtenidos hasta ahora auguran la disponibilidad del etiquetador en un plazo de tiempo razonable.

---

2. Proyecto *CRATER (Corpus Resources and Terminology Extraction, MLAP93-20)*, correspondiente al *Fifth Action Plan for the Improvement of Information Transfer between Languages.* En este proyecto colaboran además el *Laboratorio de Lingüística Informática* de la *Universidad Autónoma de Madrid*, la *Universidad de Lancaster (UK)*, *IBM-Francia* y *Computers, Communications and Visions* [Nieto 94].

En [Brent 93] se explica una forma de atacar los errores en los resultados de los marcos de subcategorización. Básicamente se trata de hacer un filtrado estadístico basado en la hipótesis de considerar correcto aquello que se repita un razonable número de veces para un verbo, teniendo en cuenta, por supuesto, el número total de apariciones del verbo. Aquellos casos que no superen unos determinados umbrales (variables para cada caso) se considerarán debidos a una de estas dos causas :

- Apariciones circunstanciales que no implican la subcategorización del verbo para ese marco. Un ejemplo típico es los complementos de lugar :

  `Me senté con cuidado.`

  Sería erróneo subcategorizar el verbo `sentar` con un marco de complemento preposicional parametrizado por `con` . En cambio sí tiene sentido reflejar la reflexividad que a menudo le acompaña. Cabe dentro de la lógica, pues, pensar que encontraremos `sentar` frecuentemente en forma reflexiva, y que, sin embargo, aparecerá acompañado por `con` tan sólo *circunstancialmente*.

- Simplemente errores del sistema :
  - En el etiquetador estocástico, cuyo carácter probabilístico ya se ha comentado.
  - En las gramáticas.

SOAMAS no incluye todavía un módulo que realice estas funciones, si bien su desarrollo está en preparación.

2.2 Criterios de diseño de las gramáticas

SOAMAS es un sistema de procesamiento de lenguaje natural basado en conocimiento sintáctico. En concreto, se han desarrollado tres gramáticas. Brevemente, la primera de ellas se encarga de analizar el grupo verbal, identificando verbos principales y auxiliares, así como posibles preposiciones y conjunciones entre ambos y pronombre proclíticos y enclíticos. La segunda se encarga de reconocer diferentes tipos de sintagmas : sintagmas nominales, adjetivos y preposicionales. Finalmente, la tercera gramática se apoya en estas dos, y describe la estructura de los complementos verbales en español. Esta última es la gramática de los marcos de subcategorización.

Es importante destacar que, en ningún momento, se ha pretendido diseñar o implementar una gramática o un analizador sintáctico del español. El enfoque para estudios de este tipo es diferente. El objetivo no es el análisis. Se pretende reconocer estructuras verbales y sintagmáticas, para posteriormente derivar de ellas los posibles marcos de subcategorización. Para ello, un analizador sintáctico no es útil, por diversas razones que se describen a continuación.

En general, un analizador sintáctico proporciona, para una oración, sus posibles descripciones estructurales dadas por la gramática. La gramática del español es indudablemente muy

compleja. Esto produce que sean varias y no una las posibles descripciones estructurales. Es más, es posible que frases correctas pero complicadas sobrepasen la capacidad del analizador implementado.

La información estructural a nivel *sintáctico* no es útil para *SOAMAS*, al que le interesa información de carácter *sintagmático*. Nos interesa conocer *qué sintagmas aparecen en el texto*. Las gramáticas de *SOAMAS* se han diseñado teniendo en cuenta este aspecto. Por otro lado, se refieren únicamente al grupo verbal y sus complementos. Consideran, por tanto, sólo una parte de la oración.

Ya se ha comentado en el apartado 1.3 los problemas que implica el punto de vista de Brent: hay muchos marcos para los que simplemente no hay *pistas* fiables ([Manning 93]). Brent se ve obligado a ceñirse a unos pocos marcos para los que sí existen dichas pistas. El diseño se las gramáticas se inspira en el enfoque propuesto por [Manning 93]. El criterio fundamental ha sido mantenerse en un compromiso entre fiabilidad y exactitud por un lado (lo que implica considerar un número reducido de casos que ofrezca pocas dudas) y relevancia (opuesto a lo anterior ya que se persigue reflejar la realidad, que es indudablemente compleja). Las gramáticas describen, pues, estructuras relativamente sencillas, pero lo suficientemente variadas como para permitir el estudio de un amplio número de marcos de subcategorización.

Se ha utilizado como referencia fundamental el estudio [Hallebeek 92], del que se tomaron numerosas ideas. En el diseño de las gramáticas ha sido necesario tomar numerosas decisiones de compromiso. Las gramáticas se han implementado usando herramientas que permiten una descripción declarativa, así como una cómoda modificación que no interacciona con el resto del sistema, pensando en un ulterior refinado de las mismas por parte de un lingüista.

3 Aspectos lingüísticos

3.1 Proposiciones simples, coordinación y subordinación

SOAMAS está orientado a la oración simple, no a la oración compuesta; no se contemplan fenómenos de coordinación o subordinación verbal. Hay varias razones para esto.

En primer lugar, no debe perderse de vista lo que ya se explicó en el apartado 2.2. Se argumentaba entonces que no se había pretendido en ningún momento desarrollar una gramática o un analizador sintáctico del español. El objetivo de *SOAMAS* es los marcos de subcategorización, y estos condicionan que el análisis se realice con la amplitud de la oración simple. Esto es debido a que los marcos de subcategorización hacen referencia a los diferentes *argumentos* que acepta un verbo. El fenómeno de *parataxis* (yuxtaposición y conjunción coordinante propiamente dicha) es una relación entre proposiciones con la misma categoría gramatical y que desempeñan la misma función en la oración compleja. Los verbos se analizan individualmente,

prescindiendo de la posibles relaciones entre ellos[1].

Más complejo es el tema de la *hipotaxis*, esto es, oraciones compuestas donde aparece una relación de subordinación entre una o varias proposiciones. La proposición o proposiciones subordinadas dependen de una principal, y pueden desempeñar en esta última una de las funciones de complemento que el verbo en cuestión admita. *SOAMAS* no es capaz de analizar correctamente la hipotaxis, a pesar de que puede ser de interés en la búsqueda de marcos de subcategorización. Conocer entre qué proposiciones se establece la relación de subordinación y qué papel (principal o subordinado) juega cada una precisaría de un dramático aumento en la complejidad del sistema.

Sin embargo, sí se analizan las proposiciones subordinadas aisladamente buscando los marcos internos.

3.2 ¿Dónde buscar los complementos?

Una vez delimitado el problema a los verbos y sus marcos respectivos en forma aislada, surge la cuestión de dónde buscar los complementos. En [Hallebeek 92; p.31-34] se proponen las siguientes estructuras para la oración simple en español:

```
(SU) (ADV) NÚCLEO (COMPL) (ADV)

(ADV) NÚCLEO (ADV) SU (COMPL) (ADV)

(ADV) NÚCLEO COMPL (ADV) SU (ADV)
```

No se ha indicado que los complementos adverbiales también pueden aparecer precediendo al sujeto o a cualquiera de los complementos que funcionen como argumentos, en aras de una mayor claridad en el esquema.

En las tres estructuras, los complementos (opcionales en algunos casos) van detrás del verbo. Esto constituye la base de la búsqueda: sólo se analiza detrás de los verbos. Delante, como se aprecia en la figura, sólo encontraremos ocasionalmente el sujeto o complementos adverbiales, ambos carentes de interés para nuestros efectos. Detrás de los verbos, sin embargo, no sólo aparecen complementos, sino que también puede aparecer el sujeto (amén de los omnipresentes complementos adverbiales).

Según Gutiérrez Araus en [Gutiérrez 78; p. 32], el 72,6% de las oraciones que encontró en el análisis de un corpus de María y d'Ors el sujeto precede al verbo. En el resto, el 27,4%, iba detrás. Estos datos se refieren, naturalemente, a los oraciones con sujeto explícito, que suponían un 65,7% del total. De ello podemos concluir que encontraremos un sujeto explícito detrás del verbo en aproximadamente un 18% de los casos. En estos casos, potencialmente erróneos, todavía es posible descubrir que se trata de un sujeto:

1. Excepto las de auxiliariedad.

- El sujeto está constituido por un sintagma nominal. Puede confundirse por tanto con un complemento directo. Si el verbo está acompañado por un pronombre clítico en función de acusativo, no cabe ya la confusión. La repetición de la función de complemento directo por un pronombre clítico es correcta, pero su aparición es muy infrecuente, ya que se utiliza con fines estilísticos.
- Incluso si no hay un pronombre en acusativo, la estructura sintáctica puede revelar la presencia del sujeto. La concatenación de dos complementos directos no es correcta en español, y por tanto podemos deducir que uno de ellos está realizando la función de sujeto.

Hay otros fenómenos en español que producen la inversión del orden normal de la oración, que generalemente se utilizan como recurso estilístico para la enfatización, y que aparecen de forma muy excepcional. El margen de error que queda es suficientemente bajo como para que la estrategia propuesta constituya una forma sólida de buscar marcos de subcategorización. Al tratamiento que se dará a estos errores esporádicos se hizo ya referencia en el apartado 2.1.

3.3 La gramática de los marcos

[Hallebeek 92; p. 113-133] señala seis tipo diferentes de complementos: atributo, sujeto atributo, objeto atributo, complemento directo, complemento indirecto y complemento preposicional[1]. Tan sólo unos comentarios al respecto.

Al igual que Hallebeek, consideramos como copulativos sólo unos pocos verbos: `ser`, `estar`, `parecer` y `semejar`. Todos ellos han de aparecer obligatoriamente acompañados por un atributo. Casos como `hacerse`, `ponerse`, `quedarse` y `seguir` no van a ser considerados como copulativos, sino como intransitivos o transitivos según el caso, con el predicado nominal en función de sujeto atributo u objeto atributo respectivamente.

El sujeto atributo y el objeto atributo son sintácticamente indistinguibles, ya que la diferencia entre ambos radica en aquello de lo que se predica: el sujeto y el complemento directo, respectivamente. La restricción de que el objeto atributo debe aparecer en compañía de un complemento directo, como se comenta más adelante al describir la *regla general* de Hallebeek, no es suficiente. Se ha optado por englobar ambos tipos de complemento en uno sólo: el *complemento atributivo*.

Hallebeek propone la siguiente regla (ver figura 2.2) que resume el orden de los complementos. Se ha excluido el atributo; al ser los verbos copulativos un número fijo no tiene interés considerarlo: cuando aparezca uno de estos verbos su comportamiento es conocido de antemano.

1. Recuérdese la diferencia entre complemento preposicional y adverbial. Cfr. 1.1.

(SA) (OA) (CI) (CD) (CP) (CI) (OA)

SA: sujeto atributo
OA: objeto atributo
CI : complemento indirecto
CD: complemento directo
CP: complemento preposicional

FIGURA 3.1 La regla general de Hallebeek sobre los complementos

Hay además unas reglas adicionales restrictivas:

- El complemento preposicional precede siempre al complemento indirecto.
- El objeto atributo únicamente puede aparecer en compañía de un complemento directo.
- El complemento preposicional no puede aparecer a la vez que un complemento directo y uno indirecto.
- Evidentemente, un complemento sólo puede aparecer con aquellos verbos que lo permitan.

La gramática de los marcos que se ha desarrollado nos permite pasar del nivel sintagmático (los sintagmas que constituyen la oración) al nivel de los marcos de subcategorización. En ella se utiliza información de muy diversos tipos:

- La regla general de Hallebeek (ver figura 2.2).
- Información del análisis del grupo verbal.
- Información del analisis de los sintagmas que siguen al verbo principal.
- Reglas restrictivas, algunas ya mencionadas.

La regla general de Hallebeek se refiere al orden que deben guardar entre sí los complementos, *pero desde el punto de vista de los complementos*. Basándose en ello y en la correspondencia en sintagmas de dichos complementos, se preparó una gramática de las diferentes combinaciones de sintagmas que pueden encontrarse detrás de un verbo. Tomemos como ejemplo un complemento atributivo seguido de uno directo: `CA_CD`. A nivel sintagmático, su reflejo es un sintagma adjetivo seguido de bien un sintagma nominal o un sintagma preposicional introducido por a: `S_ADJ S_NOM` o bien `S_ADJ S_Pa`.

Es frecuente la aparición de dos complementos en la misma oración. Se ha puesto como límite la aparición simultánea de tres complementos, si bien es posible que no se obtenga apenas información de estos casos triples.

Tenemos pues una gramática con todas las posibles combinaciones de sintagmas que van a permitirse. Toda aquella combinación que no se ajuste a esta gramática será rechazada. Esto no quiere decir, ni mucho menos, que no sea correcta en español; es más, se presupone que el cor-

pus no tiene errores gramaticales. Significa que alguna de las simplificaciones que se han introducido en *SOAMAS* afecta a esa oración en concreto: estructura demasiado compleja, subordinación, realizaciones de complementos no contempladas (p.ej., la aposición del sintagma nominal...), etc.

Finalmente es necesario tener en cuenta la información del análisis del grupo verbal en lo referente a pronombres clíticos.

3.4 Marcos simples y compuestos

Se propone un total de 25 marcos de subcategorización diferentes:

- Marcos simples: Nulo, CD, CI, CA, CP, Reflexividad, impersonalidad y función de pasiva refleja.
- Marcos compuestos: CA_CD, CA_CI, CI_CD, CD_CI, CD_CP, CD_CA, CP_CI, CA_CD_CP, CD_CI_CA, CA_CI_CD, CA_CD_CI, CI_CD_CA, CD_CP_CA, CA_CP_CI.
- Auxiliariedad: verbo auxiliar seguido de infinitivo, gerundio o participio.

Los nombres son auto-explicativos. Sí es necesario comentar la diferencia existente entre marcos simples y marcos compuestos.

Los marcos simples son los que nos proporcionan la *tipología verbal*: si un verbo admite complemento directo, indirecto... Los marcos compuestos tienen un carácter diferente: buscamos información sobre el orden que siguen los complementos para un verbo en concreto y en español en general. En otras palabras, buscamos respuesta a preguntas del tipo de cuál de las siguientes construcciones es más frecuente:

```
Entregué una carta a mi novia.

Entregué a mi novia una carta.
```

"Suena" mejor la primera versón. ¿Se trata del caso particular del verbo `entregar`? ¿Tiene acaso un carácter general? También nos interesa información de tipo cuantitativo, aunque sea con un carácter aproximado.

Consecuencia de ello es que cuando un verbo funciona con un determinado marco compuesto, se anota dicha ocurrencia en los marcos simples correspondientes. En efecto, si encontramos un marco `CD_CI`, es necesario constatar (además, claro, de dicho `CD_CI`), que hay un complemento directo y un complemento indirecto (marcos simples `CD` y `CI` respectivamente).

4 Notas sobre la implementación

SOAMAS ha sido implementado en lenguaje *C* sobre un entorno de programación *UNIX*. El sistema se haya disponible para *SunOs 4.1.2* y *Solaris 2.4* [1]. Se han utilizado herramientas típicas como *sed* y *make*; también se han usado otras pertenecientes al proyecto *GNU*[2], como el compilador *gcc* y el generador de analizadores *bison*.

Los analizadores para las gramáticas se han implementado usando el citado *bison*. Se trata de un *generador de compiladores*: produce el código *C* que corresponde al analizador de una gramática que se le proporcione. En este caso, se han desarrollado sendos analizadores para las tres gramáticas diseñadas: grupo verbal, sintagmas y marcos de subcategorización.

El sistema puede utilizarse como si fuera una función de biblioteca, de forma que pueda ser incluido dentro de otros desarrollos. Así mismo se ha implementado una interfaz más cómoda que permite un acceso directo como otro comando más.

La presentación de resultados es altamente estructurada, pensando en el citado procesado estadístico a posteriori por medio de un software basado en *awk* o *perl*.

```
* voy a entregar las instancias de la matr'icula a la secretaria del departamento
#T ------Infinitivo------ N D I A P R IMP PRF ---MC--- --PREP-- ---LOCUCION--- C V S M
&P        entregar        - x x - - -  -   -   CD_CI      -           -         - - - -
&AI          ir           - - - - - -  -   -     -        a           -
```

FIGURE 4.1 Presentación de resultados

La figura 4.1 presenta un ejemplo del formato de los resultados de *SOAMAS*, correspondiente al análisis de una oración. Se incluye el entorno del verbo encontrado, amén de, por supuesto, toda la información de subcategorización encontrada:

- Tipo de verbo:

  Debajo del código `T` se da el tipo de verbo que se ha encontrado. Los valores posibles son `P` (verbo principal), `AI` (verbo auxiliar seguido de un infinitivo), `AG` (verbo auxiliar seguido de un gerundio) y `AP` (verbo auxiliar seguido de un participio).

1. *UNIX* es una marca regristrada de *Bell Laboratories*. *SunOs* y *Solaris* son marcas registradas de *SunMicrosystems*.
2. *GNU* es un marca registrada de la *Free Software Foundation*.

- Códigos de marcos simples:

  N (nulo), D (complemento directo), I (complemento indirecto), A (complemento atibuto), P (complemento preposicional), R (reflexividad), IMP (impersonalidad), PRF (pasiva refleja).

- Códigos de marcos complejos:

  Debajo del código MC aparecen las abreviaturas de los marcos, según se explicaron en el apartado 3.4.

- Códigos de los errores:

  C (error en la codificación del conjunto palabra y etiqueta), V (error en el análisis del verbo y sus enclíticos), S (error en el análisis sintagmático) y M (error en los marcos).

5 Conclusiones

Este trabajo constituye un primer paso en el proceso de identificación de los marcos de subcategorización de los verbos españoles. El sistema desarrollado toma una oración etiquetada morfosintácticamente y trata de reconocer parcialmente la subestructura sintagmática asociada al sintagma verbal.

La utilidad de *SOAMAS* en su estado actual es limitada en dos sentidos:

1. Hasta el presente, hemos carecido de corpus etiquetados de suficiente extensión como para concluir resultados fiables. En este sentido trabajamos, fundamentalmente en colaboración con el *Laboratorio de Lingüística Informática* de la *Universidad Autónoma de Madrid*, para el desarrollo de recursos léxicos y herramientas (analizadores morfológicos y etiquetadores tanto semiautomáticos como estocásticos).

2. Por carecer aún de corpus etiquetados suficientemente extensos, hemos relegado a una segunda fase el análisis estadístico de los resultados, tarea a la que nos enfrentamos también en la actualidad. Este análisis es imprescindible además para poder evaluar con un mínimo de rigor la utilidad de *SOAMAS*.

En todo caso, la solidez de la base lingüística de *SOAMAS*, construida a partir del trabajo de formalización sintáctica del español llevado a cabo por Hallebeek, y las pruebas realizadas sobre los textos etiquetados manualmente por Aurora Martín, nos animan a continuar esta línea de trabajo; una línea cuyo objetivo inmediato es la consolidación de una plataforma léxica de propósito general para aplicaciones de procesamiento automático de textos en lengua española.

Referencias

- **[Basili *et al.* 93]** Basili, Roberto; Pazienza, María Teresa y Verlardi, Paola. *Semi-automatic Extraction of Linguistic Information for Syntactic Disambiguation.* Applied Artificial Intelligence, vol 7, p. 339-364. 1993.
- **[Brent 91]** Brent, Michael R. *Automatic Acquisition of Subcategorization Frames from Untagged Text.* Proceedings of the 29th Annual Meeting of the ACL, p.209-214. 1991.
- **[Brent 93]** Brent, Michael R. *From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax.* Computational Linguistics, vol-19, number 2, Julio 1993.
- **[Church *et al.* 90]** Church, Kenneth Ward; Hanks, Patrick. *Word Association Norms, Mutual Information and Lexicography.* Computational Linguistics, vol-16, number 1, Marzo, 1990.
- **[Cutting *et al.* 92]** Cutting, D.; Kupiec, J.; Pedersen, P.; Sibun, P.; *A Practical Part-of-Speech Tagger.* Proceedings of the Third Conference on Applied Natural Language Processing. Trento, Italia. Abril 1992.
- **[Gutiérrez 78]** Gutiérrez Araus, M.L. *Las estructuras sintácticas del español Actual.* SGEL, 1978. Madrid.
- **[Hallebeek 92]** Hallebeek, Jos. *A formal approach to Spanish syntax.* Ed. Rodopi. Amsterdam-Atlanta GA. 1992.
- **[Kupiec 92]** Kupiec, Julian M. *Robust Part-of-Speech Tagging Using a Hidden Markov Model.* Computer Speech and Language. vol 6. p. 225-242.
- **[Manning 93]** Manning, Christopher D. *Automatic Acquisition of a Large Subcategorization Dictionary from Corpora.* ACL. Mayo 1993.
- **[Martín 94]** Martín de Santa Olalla Sánchez, Aurora. *Una propuesta de codificación morfosintáctica para corpus de referencia en lengua española.* Tesis Doctoral. Universidad Autónoma de Madrid, 1994. Madrid.
- **[Nieto 94]** Nieto, Amalio F. *CRATER: UPM Progress for the Period April-September 1994* . Internal Project Document. DIT, UPM. Septiembre 1994.
- **[Rouvret *et al.* 80]** Rouvret y Vergnaud. *Specifying Reference to the Subject.* Linguistic Inquiry. vol 11-1. 1980.