Procesamiento del Lenguaje Natural, num. 29 (2002), pp.31-38 recibido 10-05-2002; aceptado 02-07-2002

Usando la co-composicionalidad para el aprendizaje de la subcategorización sintáctico-semántica *

Pablo Gamallo Otero **Alexandre Agustini** **Gabriel P. Lopes**
Dept. de Informática, Univ. Nova de Lisboa, Portugal
gamallo,aagustini,gpl@di.fct.unl.pt

Resumen: El análisis sintáctico precisa de léxicos expandidos con información sobre subcategorización. Este artículo describe un método no supervisado capaz de adquirir restricciones sintácticas y semánticas a partir de corpora restrictos. Se dará especial atención al papel de la co-composición en el proceso de aprendizaje de este tipo de información.
Palabras clave: Extracción de información semántica, Adquisición de restricciones de selección, Corrección del análisis sintáctico, Léxico Generativo

Abstract: Natural language parsing requires extensive lexicons containing subcategorisation information for specific sublanguages. This paper describes an unsupervised method for acquiring both syntactic and semantic subcategorisation restrictions from corpora. Special attention will be paid to the role of co-composition in the acquisition strategy.
Keywords: Semantic Information Extraction, Selection Restrictions Acquisition, Parsing Improvement, Generative Lexicon

1 *Introducción*

Las gramáticas lexicalistas proyectan en las estructuras sintácticas los patrones de subcategorización codificados en el léxico. Estas gramáticas se sirven de léxicos con información específica sobre subcategorización a fin de restringir el análisis sintáctico. Los analizadores sintácticos necesitan esta información para reducir el número de análisis posibles y, con ello, resolver la ambigüedad estructural. En los últimos años, han ido apareciendo diferentes métodos para adquirir patrones de subcategorización a partir de corpora. Algunos de ellos tienen por objeto inducir la subcategorización sintáctica a partir de textos anotados (Brent, 1993; Briscoe y Carrol, 1997; Marques, 2000). Desgraciadamente, el recurso a la información sintáctica no es suficiente para resolver la ambigüedad estructural. Tomemos como ejemplo los dos siguientes análisis:

(1) [pelar [$_{SN}$ la patata] [$_{SP}$ con un cuchillo]]
(2) [pelar [$_{SN}$ [$_{SN}$ la patata] [$_{SP}$ con una mancha]]]

Los criterios puramente sintácticos no son suficientes para conseguir unir *con_SP* al verbo *pelar* en (1) y al SN *la patata* en (2). Supongamos que la dependencia entre el SP *con un cuchillo* y el verbo *pelar* es sugerida por el hecho de que este último subcategoriza expressiones de tipo *con_SP*. Esto no explica, sin embargo, el análisis (2), donde el SP *con una mancha* aparecerá ligado directamente al sintagma anterior: *la patata*. Para conseguir un análisis correcto, necesitamos recurrir a nuestro conociemento del mundo, en particular, a nuestros conocimientos acerca de la acción de pelar, el uso de cuchillos y las propiedades de las patatas. En general, sabemos que los cuchillos sirven para pelar y que las patatas pueden tener manchas en la piel. De esta manera, para llegar a un análisis correcto, el analizador debe extraer del léxico, no sólo información acerca de la subcategorización sintáctica, sino también acerca de las preferencias semánticas.

Existen otros métodos que tienen como objetivo extraer restricciones de selección usando estrategias de adquisición supervisadas. El algoritmo de aprendizaje requiere que en el corpus de muestra estén anotados con etiquetas semánticas los nombres que aparecen en dependencias de tipo verbo-nombre, nombre-verbo, o verbo-prep-nombre. En algunos casos, la anotación es manual, y en otros con base en diccionarios y tesauri construidos manualmente (Resnik, 1997; Framis, 1995).

También se están desarrollando estrate-

* Work supported by Fundação para a Ciência e a Tecnologia, Ministério da Ciência e a Tecnologia, Portugal.

gias no supervisadas para la adquisición de selecciones de restricción. Estas no dependen de un muestreo textual previamente anotado con etiquetas semánticas (Sekine et al., 1992; Dagan, Lee, y Pereira, 1998; Grishman y Sterling, 1994). De acuerdo con la terminología de Grefensetette, tales estrategias pueden clasificarse como "pobres en conocimiento" (Grefenstette, 1994). Las preferencias semánticas son inducidas usando apenas coocurrencias de palabras, en particular se acostumbra a utilizar diferentes tipos de medidas de semejanza con el fin de identificar aquellas palabras que coocurren en las mismas dependencias sintácticas. Supongamos, por ejemplo, que el verbo *ratificar* aparece frecuentemente con el nombre *organización* en la posición de sujeto gramatical. Supongamos además que este nombre tiene una distribución similar a la de otros nombres con significado "animado", tales como *secretario* y *ayuntamiento*. Podemos inferir que *ratificar* no sólo selecciona *organización*, sino también aquellas palabras que se le asemejan. Esto parece correcto. Sin embargo, supongamos ahora que *organización* también tiene una alta frecuencia en contextos lingüísticos como *la organización del equipo falló de nuevo* o *nadie trabajó en la organización del espectáculo*. En estos contextos, *organización* vehicula un significado diferente al anterior: aquí designa un tipo particular de proceso o evento. Evidentemente, las palabras *secretario* y *ayuntamiento* no aparecen en estos nuevos contextos, ya que sólo se relacionan con *organización* con respecto a su otro uso y significado. Para solucionar este problema, vamos a proponer un método de "clustering" en que cada palabra puede aparecer incluida en diferentes "clusters", y donde cada cluster representa las restricciones semánticas impuestas por una clase de contextos de subcategorización.

Este artículo describe una metodología no supervisada cuyo objetivo es la adquisición de contextos de subcategorización sintáctico-semánticos, tanto para nombres como para verbos, y todo ello a partir de corpora parcialmente analizados. A continuaciión, expondremos las principales hipótesis de nuestra propuesta. Después, en la section 3, se presentarán los dos procesos en los que se basa el método de aprendizaje [1]. En la sección 4, mostraremos como se introduce la información aprendida en el diccionario utilizado por el analizador sintáctico. Finalmente, mediremos la precisión y cobertura de esta información en relación a una aplicación específica: la resolución de dependencias sintácticas.

2 Hipótesis Lingüísticas

Nuestro método de adquisición gira en torno a dos hipótesis básicas. En primer lugar, consideramos que la noción de subcategorización lingüística debe ser definida en base a la idea de "co-composición", tal como es definida en los trabajos de Pustejovsky (Pustejovsky, 1995). De esta manera, debemos asumir que, en una dependencia de tipo "núcleo-modificador" (*head-modifier*), no sólo el núcleo impone restricciones de selección al modificador, sino que éste también impone algún tipo de preferencia semántica al núcleo. Para una palabra determinada, intentaremos adquirir, por tanto, dos tipos de subcategorización: los modificadores con los que se combina, por un lado, y los núcleos que modifica, por otro. Veamos, por ejemplo, cual podría ser el comportamiento composicional del nombre *república* en un corpus de dominio específico. Por un lado, esta palabra puede aparecer en la posición de núcleo en dependencias tales como *república de Irlanda, república de Portugal*, etc. Por otro lado, también aparece en la posición de modificador en dependencias como *presidente de la república, gobierno de la república* , etc. Dado que existen interesantes regularidades entre las palabras que coocurren con *república* en tales contextos, nuestro objetivo será el de implementar un algoritmo con capacidad para aprender dos tipos de contextos de subcategorización:

- $[\lambda x^{\uparrow}(de; república^{\downarrow}, x^{\uparrow})]$ donde la preposición *de* representa una relación binaria entre la palabra *república*, que desempeña el papel de "núcleo" (papel anotado por medio de la flecha "$\downarrow$"), y todas aquellas palabras que pueden llegan a ser sus "modificadores" (usamos "$\uparrow$" para el papel de modificador). Desde el punto de vista semántico, este contexto de subcategorización requiere complementos que designen naciones o estados específicos, ya que sólo las naciones y estados pueden ser repúblicas.

[1] Las experiencias descritas en este artículo fueron realizadas sobre 1,5 million de palabras pertenecientes al corpus *P.G.R.* (*Procuradoria Geral da República*), que contiene documentos en portugués en torno a temas de naturaleza legal-administrativa.

• $[\lambda x^{\downarrow}(de; x^{\downarrow}, república^{\uparrow})]$ este es un contexto de subcategorización que requiere un núcleo que designe alguna organización, institución, cargo o función de la república.

Nuestra noción de subcategorización engloba por tanto no sólo preferencias sintácticas sino también semánticas.

La segunda hipótesis en la que basamos nuestro método se refiere al proceso de construcción de clases de contextos de subcategorización. Postulamos, en particular, que dos contextos de subcagorización pueden ser considerados semánticamente similares si aparecen en el corpus con la misma distribución. Tomemos, por ejemplo, los siguientes contextos:

$[\lambda x^{\downarrow}(de;x^{\downarrow},república^{\uparrow})]$ $[\lambda x^{\downarrow}(of;x^{\downarrow},estado^{\uparrow})]$
$[\lambda x^{\uparrow}(de;oficina^{\downarrow},x^{\uparrow})]$ $[\lambda x^{\uparrow}(iobj\_a;incumbir^{\downarrow},x^{\uparrow})]$

Dado que todos ellos parecen compartir las mismas preferencias semánticas, tienden a tener una distribución de palabras muy similar. Las palabras que aparecen con más frecuencia en todos estos contextos representan la clase con la que se puede definir extensionalmente sus preferencias semánticas. Por ejemplo, las palabras *ministro, presidente, asamblea, ...*, que se distribuyen regularmente en estos contextos, pueden servir para construir la clase extensional con la que se describe sus restricciones de selección.

A diferencia de los métodos no supervisados más utilizados en la extracción de restricciones de selección, nuestro modelo no busca identificar palabras similares en base a la *hipótesis distribucional* de Harris. Según esta hipótesis, las palabras que coocurren en los mismos contextos de subcategorización son semánticamente similares. Tal como hemos explicado en la Introducción, esta concepción acerca de la semejanza entre palabras no parece poder tomar en cuenta el fenómeno de la polisemia. Por ello, el principal objetivo de nuestro método será el de medir la semejanza, no entre palabras, sino entre contextos de subcategorización (Faure y Nédellec, 1998; Faure, 2000).

3 Adquisición de la subcategorización

Con el fin de evaluar las dos hipótesis que acabamos de formular, fue elaborado un método de adquisición de subcategorización sintáctica y semántica a partir de un corpus parcialmente analizado. Este método se compone de dos procesos secuenciales. El primero tiene como objetivo extraer contextos posibles de subcategorización (que llamamos "contextos candidatos"), mientras que el segundo identifica, entre los contextos candidatos, aquéllos que son correctos, a fin de agruparlos en clases semánticas de subcategorización. A continuación, estos dos procesos serán definidos de una manera más precisa.

3.1 Extracción de contextos candidatos

Se etiqueta primero el texto por medio de "*tags*" morfosintácticos (Marques, 2000), y luego se analiza parcialmente (Rocio, de la Clergerie, y Lopes, 2001) en secuencias de "*chunks*" básicos (SN, SP, SV, ...). Seguidamente, por medio de heurísticas simples de tipo "asociación a la derecha", asignamos estructuras sintácticas temporales: por ejemplo, consideramos que un chunk tiende a estar relacionado sintácticamente con el chunk que se encuentra inmediatamente a su izquierda. Tomando en cuenta la primera hipótesis formulada en 2, asumimos que dos chunks ligados sintácticamente forman una dependencia binaria susceptible de dividirse en dos contextos de subcategorización. Dado que las heurísticas estructurales empleadas no toman en cuenta las dependencias a distancia, es evidente que pueden surgir varios tipos de errores sintácticos.[2] Por este motivo, los contextos de subcategorización identificados en esta primera etapa se consideran sólo contextos candidatos. Finalmente, también postulamos que el conjunto de palabras asociado a cada contexto candidato representa lo que podría ser una clase semántica. Por ejemplo, la frase

emanou de facto da lei
(*emanó de hecho de la ley*)

da lugar a las dos siguientes dependencias:

$(iobj\_de; emanar^{\downarrow}, facto^{\uparrow})$ $(de; facto^{\downarrow}, lei^{\uparrow})$

a partir de las que se pueden generar cuatro posibles contextos de subcategorización:

$[\lambda x^{\downarrow}(iobj\_de;x^{\downarrow},facto^{\uparrow})]$ $[\lambda x^{\uparrow}(iobj\_de;emanar^{\downarrow},x^{\uparrow})]$
$[\lambda x^{\downarrow}(de;x^{\downarrow},lei^{\uparrow})]$ $[\lambda x^{\uparrow}(de;facto^{\downarrow},x^{\uparrow})]$

[2] los errores pueden ser causados, no sólo por las limitaciones de estas heurísticas, sino también por las propias incorrecciones del corpus, la incompletud del diccionario, palabras incorrectamente etiquetadas, etc.

Sin embargo, estos cuatro contextos candidatos son incorrectos. El complemento preposicional `de facto` representa una locución adverbial interpolada entre el verbo y su verdadero complemento, `da lei`. Veremos más adelante cómo se adquire la información sobre subcategorización, y también cómo ésta se utiliza para reconocer las dependencias incorrectas. El algoritmo de aprendizaje se basa esencialmente en el cálculo de la semejanza entre los conjuntos de palabras asociados a los contextos de subcategorización candidatos.

3.2 Generar clases de contextos de subcategorización

De acuerdo con la segunda hipótesis formulada en la sección 2, dos contextos de subcategorización con una distribución de palabras semejante deben contener las mismas restricciones de selección. En torno a esta idea, vamos a desarrollar el proceso de generación de clases de subcategorización semántica. Este proceso consta de dos tareas: "filtración" y "*clustering*".

3.2.1 Filtración

A cada contexto de subcategorización se le asocia un conjunto de palabras con sus respectivas frecuencias, llamado "conjunto contextual". Nuestro objetivo es reunir conjuntos contextuales similares en conjuntos cada vez más grandes que representen extensionalmente clases de subcategorización semántica. El problema reside en que la mayor parte de los conjuntos contextuales contiene palabras que fueron mal etiquetadas o que provienen de estructuras sintácticas incorrectas. El objeto de la filtración es, precisamente, retirar estas palabras de los conjuntos contextuales. El proceso de filtración se realiza de la siguiente manera.

En primer lugar, a cada conjunto contextual se le asocia una lista de conjuntos similares. Cuanto mayor sea el porcentaje de palabras que comparten dos conjuntos, mayor es su grado de semejanza. Entre los diferentes coeficientes estadísticos que fueron usados para extraer listas de conjuntos semejantes, los mejores resultados se consiguieron con una versión con pesos del coeficiente Jaccard. Atribuimos pesos a las palabras en función de su dispersión (peso global) y de su frecuencia relativa (peso local) en cada contexto o conjunto contextual. La dispersión de la palabra (peso global) *disp* toma en cuenta el número de contextos asociados a una palabra dada y su frecuencia en el corpus. El peso local se calcula tomando en cuenta la frecuencia relativa fr del par palabra/contexto. El peso P de la palabra *pal* con respecto al contexto *cntx* se calcula de esta manera:

$$P(pal_i, cntx_j) = log_2(fr_{ij}) * log_2(disp_i)$$

donde

$$fr_{ij} = \frac{frecuencia\,de\,pal_i\,con\,cntx_j}{\sum_i frecuencia\,de\,pal_i\,con\,cntx_j}$$

y

$$disp_i = \frac{\sum_j frecuencia\,de\,pal_i\,con\,cntx_j}{numero\,de\,contextos\,con\,pal_i}$$

Finalmente, la medida Jaccard de semejanza SJ entre dos contextos m y n es el resultado de [3]:

$$SJ(cntx_m,cntx_n)=\frac{\sum_{comun_i}(P(cntx_m,pal_i)+P(cntx_n,pal_i))}{\sum_j(P(cntx_m,pal_j)+P(cntx_n,pal_j))}$$

Para cada conjunto contextual (y su correspondiente contexto de subcategorización), seleccionamos una lista de contextos considerados semejantes. Por cada par de conjuntos semejantes, extraemos sólo las palabras que comparten. Es decir, seleccionamos la intersección de palabras comunes de cada par de conjuntos contextuales que fueron considerados semejantes. Cada intersección representa una clase semántica homogénea, que llamamos "clase básica". Tomemos un ejemplo. En nuestro corpus, el contexto estadísticamente más parecido a $[\lambda x^{\uparrow}(de; violação^{\downarrow}, x^{\uparrow})]$ es el contexto: $[\lambda x^{\uparrow}(dobj; violar^{\downarrow}, x^{\uparrow})]$. Ambos comparten las siguientes palabras:

```
sigilo princípios preceito plano
norma lei estatuto disposto disposição
direito
```

Esta clase básica no contiene palabras como `vez, flagrantemente, obrigação, interesse`, que aparecen asociadas por error a uno de los dos contextos. El proceso de filtración consigue, por tanto, eliminar este tipo de errores. Esta clase básica puede ser considerada semánticamente homogénea pues las

[3] *comun* restringe el sumatorio a las palabras comunes a m y n.

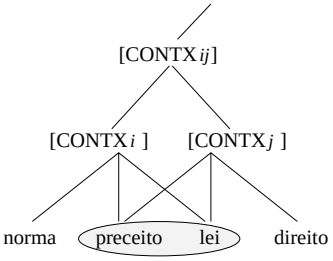

Figura 1: Proceso de clustering

palabras componentes designan todas el mismo tipo de entidades: todas se refieren a documentos legales. Una vez que las clases básicas han sido generadas, entra en juego el algoritmo de *clustering*, que toma como punto de partida estas clases con el fin de producir otras más generales.

3.2.2 Clustering conceptual

Usamos un proceso de clustering aglomerativo (*botton-up*), para ir agregando progresivamente las clases básicas en nuevas clases derivadas. Los detalles de este proceso aparecen en (Gamallo, Agustini, y Lopes, 2001). La Figura 1 ilustra dos clases básicas generadas por dos pares de contextos de subcategorización similares. $[CONTX_i]$ representa un par de contextos de subcategorización similares. Estos dos contextos semejantes tienen en común la clase básica: `preceito, lei, norma` (*precepto, ley, norma*). Por otro lado, $[CONTX_j]$ representa otro par de contextos similares, que comparten las palabras `preceito, lei, direito` (*precepto, ley, derecho*). Ambas clases básicas fueron generadas por el proceso de filtración descrito en la sección anterior. El proceso de clustering las junta en una clase más general constituida por `preceito, lei, norma, direito`. Al mismo tiempo, los dos pares de contextos, $[CONTX_i]$ y $[CONTX_j]$, también se reúnen en un nuevo conjunto de contextos: $[CONTX_{ij}]$. Este proceso de generalización nos permite inducir información sintáctica que no aparece explícitamente en el corpus. En particular, inducimos que la palabra `norma` puede aparecer en los contextos de subcategorización representados por $[CONTX_j]$, y también que `direito` puede ocurrir en los contextos representados por $[CONTX_i]$.

En el proceso de clustering extendemos, por tanto, una clase básica (o preferencia semántica) en dos direcciones: aumentamos el número de palabras que constituyen la clase (o preferencia), al mismo tiempo que incrementamos el número de contextos que imponen esa clase (o preferencia).

Tabla 1: Clases de la palabra `trabalho`

Cluster_1	`contrato execução exercício prazo processo procedimento trabalho` *(contrato ejecución ejercicio plazo proceso procedimiento trabajo)*
Cluster_2	`contrato exercício prestação recurso serviço trabalho` *(contrato ejercicio prestación recurso servicio trabajo)*
Cluster_3	`actividade atribuição cargo exercício função lugar trabalho` *(actividad attribución cargo ejercicio función puesto trabajo)*

3.2.3 Representación de las palabras polisémicas

Las palabras polisémicas aparecen en diferentes clusters. Por ejemplo, la Tabla 1 sitúa la palabra `trabalho` (*trabajo*) en tres clases diferentes. Cluster_1 contiene palabras que designan objetos temporales. Estas palabras aparecen, por tanto, en contextos de subcategorización que exigen argumentos temporales: e.g., $[\lambda x^{\uparrow}(de; suspensão^{\downarrow})]$ y $[\lambda x^{\downarrow}(em; x^{\downarrow}, curso^{\uparrow})]$. Cluster_2 agrega palabras que describen el resultado de una acción. Este significado es relevante en contextos como: $[\lambda x^{\uparrow}(iobj\_por; receber^{\downarrow}, x^{\uparrow})]$. En efecto, el motivo por el que se recibe dinero no es la acción de trabajar, sino el producto realizado a través del trabajo. Finalmente, Cluster_3 ilustra el significado "tarea" o "función". Este sentido se activa en contextos tales como: $[\lambda x^{\downarrow}(de; inspector^{\downarrow})]$, o $[\lambda x^{\downarrow}(dobj; desempenhar^{\downarrow}, x^{\uparrow})]$.

4 Applicación y evaluación

Las clases adquiridas en la etapa de aprendizaje serán utilizadas por el analizador sintáctico. Para ello, enriquecemos primero el léxico con información sobre subcategorización sintáctico-semántica. Una vez actualizado el diccionario léxico, realizamos un segundo ciclo de análisis sintáctica, con el fin de verificar y corregir las dependencias que fueron identificadas en el primer ciclo.

4.1 Actualización del léxico

La Tabla 2 muestra tres entradas léxicas enriquecidas con información sobre subcategorización sintáctico-semántica. Cada entrada contiene dos listas de información: los

Tabla 2: Dictionary entries

- **abono**
  - $[\lambda x^{\downarrow}(de; x^{\downarrow}, abono^{\uparrow})]$ =

{aplicação caso fixação montante pagamento título}

  - $[\lambda x^{\uparrow}(de; abono^{\downarrow}, x^{\uparrow})]$ =

{ajuda despesa pensão quantia remuneração subsídio suplemento valor vencimento}

  - $[\lambda x^{\downarrow}(iobj\_de; x^{\downarrow}, abono^{\uparrow})]$ =

{conceder conter definir determinar fixar manter prever}

- **emanar**
  - $[\lambda x^{\uparrow}(iobj\_de; emanar^{\downarrow}, x^{\uparrow})]$ =

{alínea artigo código decreto diploma disposição estatuto legislação lei norma regulamento}

  - $[\lambda x^{\uparrow}(iobj\_de; emanar^{\downarrow}, x^{\uparrow})]$ =

{administração autoridade comissão conselho direcção estado governo ministro tribunal órgão}

- **presidente**
  - $[\lambda x^{\uparrow}(de; presidente^{\downarrow}, x^{\uparrow})]$ =

{assembleia câmara comisão conselho direcção estado empresa gestão instituto região república secção tribunal órgão}

  - $[\lambda x^{\downarrow}(de; x^{\downarrow}, presidente^{\uparrow})]$ =

{cargo categoria função lugar remuneração vencimento}

contextos de subcategorización y las clases de palabras requeridas por cada uno de los contextos. Como ya hemos dicho anteriormente, tales clases representan, en extensión, las preferencias semánticas de los contextos de subcategorización a los que están asociados. Veamos la información que nuestro sistema consiguió adquirir sobre el verbo `emanar` (ver Tabla 2). Este verbo subcategoriza sintácticamente dos tipos de "de-modificadores": uno de ellos requiere semánticamente palabras que designan documentos legales (`emana da lei`); el otro selecciona palabras que denotan instituciones (`emana da autoridade`). La información sobre este tipo de restricciones semánticas es apropiada para corregir las dependencias que fueron erróneamente propuestas por el analizador para frases como `emanou de facto da lei`. Dado que la palabra `facto` no pertenece a la clase semántica que el verbo requiere en la posición "de-modificador", la dependencia entre `emanou` y `de facto` no se puede confirmar. Posteriormente, en un nuevo ciclo se intentará ligar el verbo con el siguiente posible complemento: `da lei`.

En cuanto a los nombres `abono` y `presidente`, éstos subcategorizan no sólo modificadores sino también diferentes tipos de núcleo. Por ejemplo, `abono` selecciona núcleos nominales que lo requieren como "de-modificador": `fixação` (`fixação do abono`), así como núcleos verbales que lo requieren como modificador de objeto directo: `fixar o abono`.

4.2 Algoritmo para la resolución de dependencias

La información ya introducida en el léxico se utiliza, a su vez, para verificar si los contextos candidatos previamente extraidos por el analizador son dependencias correctas. El grado de eficacia en esta tarea de resolución puede ayudar a evaluar la robustez de nuestra estrategia de aprendizaje.

Con el fin de mejorar el análisis sintáctico, un "sistema de diagnóstico" (Rocio, de la Clergerie, y Lopes, 2001) recibe en entrada las dependencias candidatas propuestas en el primer ciclo, comprueba si son correctas con base en la información sobre subcategorización, y propone procedimientos de corrección. Tomemos, por ejemplo, la expresión `editou o artigo` (*editó el artículo*). El sistema de diagnóstico propone que la dependencia candidata $(dobj; editar^{\downarrow}, artigo^{\uparrow})$ sea verificada por el sistema. La verificación se basa en la aceptación o rechazo de la candidata. Este proceso, que se realiza tomando en cuenta la nueva información introducida en el léxico, consta de 4 subtareas:

Tarea 1a - Verificación sintáctica de `artigo`: se busca en el léxico la palabra `artigo`, y se comprueba si posee la restricción sintáctica: $[\lambda x^{\downarrow}(dobj; x^{\downarrow}, artigo^{\uparrow})]$. Si `artigo` tiene esta restricción, entonces, pasamos a la verificación semántica. En caso contrario, pasamos a la tarea 2a.

Tarea 1b - Verificación semántica de `artigo`: se comprueba la clase semántica asociada a $[\lambda x^{\downarrow}(dobj; x^{\downarrow}, artigo^{\uparrow})]$. Si la palabra `editar` pertenece a esta clase, entonces inferimos que $(dobj; editar^{\downarrow}, artigo^{\uparrow})$ es una relación binaria. La dependencia es confirmada. En caso contrario, pasamos a la tarea 2a.

Tarea 2a - Verificación sintáctica de `editar`: se busca en el léxico la palabra `editar`, y se comprueba si posee la restricción sintáctica: $[\lambda x^{\uparrow}(dobj; editar^{\downarrow}, x^{\uparrow})]$. Si `editar` tiene esta restricción, entonces, pasamos a la verificación semántica. En caso contrario, pasamos a la tarea 2a.

Tarea 2b - Verificación semántica de `editar`: se comprueba la clase semántica asociada a $[\lambda x^{\uparrow}(dobj; editar^{\downarrow}, x^{\uparrow})]$. Si la palabra `artigo` pertenece a esta clase, entonces infe-

rimos que ($dobj$; $editar^{\downarrow}$, $artigo^{\uparrow}$) es una relación binaria. La dependencia es confirmada. En caso contrario, la dependencia no puede ser confirmada.

La verificación semántica se basa en la hipótesis sobre la co-composición, introducida anteriormente (sección 2). Según esta hipótesis, dos palabras son sintácticamente dependientes sólo si una de estas dos condiciones se confirma: el núcleo requiere semánticamente al modificador, o el modificador requiere semánticamente al núcleo. Debido a las propiedades de la co-composicionalidad, basta apenas que se realize una de estas dos condiciones para poder afirmar que dos palabras son sintácticamente dependientes.

4.3 Evaluación de del procedimiento de resolución de dependencias

La Tabla 3 ilustra la evaluación manual de las correcciones propuestas por el sistema de diagnóstico. Evaluamos la precisión y la cobertura de las correcciones propuestas en torno a tres tipos de dependencias candidatas: $prep(N^{\downarrow}, N^{\uparrow})$, $dobj(V^{\downarrow}, N^{\uparrow})$, y $iobj\_prep(V^{\downarrow}, N^{\uparrow})$. Llamamos *precisión* al porcentaje de verificaciones correctas con respecto al número de correcciones efectivamente realizadas. La *cobertura* indica la proporción de dependencias candidatas que fueron efectivamente verificadas.

Tabla 3: Evaluación del proceso de resolución de tres tipos de dependencias

Dep. Candidata	Precisión (%)	Cobertura (%)
$prep(N^{\downarrow},N^{\uparrow})$	95,53	30,27
$dobj(V^{\downarrow},N^{\uparrow})$	94,44	19,44
$iobj\_prep(V^{\downarrow},N^{\uparrow})$	93,87	10,11
Total	94,61	19,94

Mientras que la precisión alcanza valores elevados (sobre 95%), la cobertura apenas logra el 20%. El motivo de esta baja cobertura se debe a que la información sobre subcategorización sólo puede ser adquirida cuando las palabras aparecen con cierta frecuencia en el corpus. Por consiguiente, la cobertura podrá ser incrementada a medida que la talla del corpus aumente.

5 Trabajo futuro

Dado que el método que hemos presentado todavía no ofrece propuestas para las dependencias a distancia, no es posible compararlo con los trabajos ya existentes que toman en cuenta este tipo de dependencias (Hindle y Rooth, 1993; Brill y Resnik, 1994). En nuestro enfoque, la resolución de dependencias a distancia sólo será posible una vez que se agoten las correcciones sobre las dependencias inmediatas. Estamos actualmente trabajando en la elaboración de nuevos ciclos de análisis centrados en la verificación y corrección de dependencias lejanas. Tomemos de nuevo la frase `emanou de facto da lei`. Como ya fue dicho anteriormente, el sistema de diagnóstico confirmó que el SP `de facto` no puede ser ligado al verbo `emanou`. En un nuevo ciclo de análisis, el sistema deberá verificar si el segundo SP `da lei` puede ser directamente ligado al verbo. El sistema ejecutará n-ciclos de anánalis, hasta agotar el número de candidatos posibles. Al final del proceso, el parser deberá proponer aquellas dependencias que mejor se adapten a la información disponible en el léxico.

Bibliografía

Brent, Michael. 1993. From grammar to lexicon: unsupervised learning of lexical syntax. *Computational Linguistics*, 19(3):243–262.

Brill, Eric y Philip Resnik. 1994. A rule-based approach to prepositional phrase attachment disambiguation. En *COLING*.

Briscoe, Ted y John Carrol. 1997. Automatic extraction of subcategorization from corpora. En *5th Conference on Applied Natural Languague Processing (ANCP97)*, Washington, DC, USA.

Dagan, Ido, Lillian Lee, y Fernando Pereira. 1998. Similarity-based methods of word coocurrence probabilities. *Machine Learning*, 43.

Faure, David. 2000. *Conception de méthode d'aprentissage symbolique et automatique pour l'acquisition de cadres de sous-catégorisation de verbes et de connaissances sémantiques à partir de textes : le système ASIUM*. Ph.D. tesis, Université Paris XI Orsay, Paris, France.

Faure, David y Claire Nédellec. 1998. Asium: Learning subcategorization frames and

restrictions of selection. En *ECML98, Workshop on Text Mining.*

Framis, Francesc Ribas. 1995. On learning more appropriate selectional restrictions. En *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, Dublin.

Gamallo, Pablo, Alexandre Agustini, y Gabriel P. Lopes. 2001. Selection restrictions acquisition from corpora. En *10th Portuguese Conference on Artificial Intelligence (EPIA'01)*, páginas 30–43, Porto, Portugal. LNAI, Springer-Verlag.

Grefenstette, Gregory. 1994. *Explorations in Automatic Thesaurus Discovery.* Kluwer Academic Publishers, USA.

Grishman, Ralph y John Sterling. 1994. Generalizing automatically generated selectional patterns. En *Proceedings of the 15th International on Computational Linguistics (COLING-94).*

Hindle, Donald y Mats Rooth. 1993. Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1):103–120.

Marques, Nuno. 2000. *Uma Metodologia para a Modelação Estatística da Subcategorização Verbal.* Ph.D. tesis, Universidade Nova de Lisboa, Lisboa, Portugal.

Pustejovsky, James. 1995. *The Generative Lexicon.* MIT Press, Cambridge.

Resnik, Philip. 1997. Selectional preference and sense disambiguation. En *ACL-SIGLEX Workshop on Tagging with Lexical Semantics*, Washinton DC.

Rocio, V., E. de la Clergerie, y J.G.P. Lopes. 2001. Tabulation for multi-purpose partial parsing. *Journal of Grammars*, 4(1).

Sekine, Satoshi, Jeremy Carrol, Sofia Ananiadou, y Jun'ichi Tsujii. 1992. Automatic learning for semantic collocation. En *Proceedings of the 3rd Conference on Applied Natural Language Processing*, páginas 104–110.