JESÚS PERAL *
PATRICIO MARTÍNEZ-BARCO *
RAFAEL MUÑOZ *
ANTONIO FERRÁNDEZ *
LIDIA MORENO **
MANUEL PALOMAR *
***Dpto. de Lenguajes y Sistemas Informáticos, Universidad de Alicante, España.**
****Dpto. de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, España.**

Una técnica de análisis parcial sobre textos no restringidos (SUPP) aplicada a un sistema de extracción de información (EXIT).

Resumen

En este trabajo presentamos la integración de una técnica de análisis parcial definida sobre textos no restringidos (SUPP) sobre un sistema de extracción de información (EXIT) en el dominio restringido de los textos notariales, en concreto, en las escrituras de compraventa de inmuebles. La principal contribución está centrada en la identificación de sintagmas nominales definidos, acrónimos y alias para su resolución posterior de correferentes.

1. Introducción.

Tradicionalmente, un analizador realizaba un análisis completo de una oración y sólo aceptaba aquellas oraciones que eran reconocidas por la gramática. Partíamos, pues, de un universo de discurso cerrado con una gramática completa para ese universo, y por tanto el objetivo era encontrar el mejor analizador para la gramática.

Cuando tratamos con textos no restringidos, los analizadores no pueden realizar análisis completos por falta de información léxica y por falta de reglas gramaticales. Los analizadores parciales surgen como solución a este problema. La idea que plantean estos sistemas es la de realizar un análisis de estructuras de información relevantes (generalmente pequeñas) y que puedan recuperarse con poca información sintáctica, a diferencia de un sistema de análisis completo que intenta recuperar estructuras de información grandes que requieren mucha información. Estos mecanismos de análisis parciales se están aplicando a temas concretos como la resolución de la anáfora, la extracción/recuperación de información y/o la elaboración automática de resúmenes.

En este trabajo presentaremos, en primer lugar, la técnica de análisis parcial *SUPP* (*Slot Unification Partial Parser*) según [Mar98a], como un método de análisis que a partir de un conjunto de reglas gramaticales obtenga la información relevante según estas reglas. Este método es capaz de extraer sintagmas nominales, sintagmas preposicionales y chunks[1] verbales en textos no restringidos y a su vez está capacitado para la resolución sintáctica de fenómenos lingüísticos (anáfora, elipsis y extraposición de elementos) que puedan aparecer en el texto. La evaluación se realizó mediante un corpus de entrenamiento etiquetado, LESESP, en el que no existe ambigüedad léxica [Mar98b].

En segundo lugar, aplicaremos este método de análisis a un sistema de extracción de información (EXIT) [Llop98]. El sistema SUPP es una buena herramienta para ser aplicada a las tareas de reconocimiento de entidades y resolución de correferencias en un sistema de extracción de información. Destacar que la aplicación de este método se realizará sobre un conjunto de escrituras notariales sin etiquetar, siendo un dominio semánticamente restringido.

2. Sistema SUPP.

En este apartado describiremos, en primer lugar, la gramática *SUG* (*Slot Unification Grammar*) como base del analizador parcial SUPP y a continuación el propio sistema SUPP.

2.1. Gramática SUG.

Las SUG fueron desarrolladas en [Fer97] como una extensión de las *Gramáticas de Cláusulas Definidas (Definite Clause Grammars, DCG)* [Per80] con el objetivo de ampliar las capacidades de las DCG para facilitar la resolución de manera modular de diversos problemas lingüísticos. Las SUG se

[1] Un chunk se define como una secuencia de elementos con cierto sentido sintáctico alrededor de un núcleo o cabecera [Abn97].

denominan así debido a las *estructuras de huecos* (*slot structures, EH*) generadas automáticamente por el analizador, donde se incluyen de forma automática toda la información morfológica, sintáctica y semántica necesaria para resolver problemas lingüísticos variados [Fer98].

Las SUG se definen según [Fer97] como una cuádrupla *(NT,T,P,H)*, donde *NT* es un conjunto finito de símbolos no terminales y *T* es un conjunto finito de símbolos terminales disjunto con *NT. P* son las *reglas de producción* de la gramática: un conjunto finito de pares α++>β donde $\alpha \in NT$, $\beta \in (T \cup NT)^* \cup$ *{llamadas a procedimientos}*. Por último, H son *hechos SUG*: reglas de producción que sólo tienen el primer miembro de la regla, donde α puede ser *coordination*, *juxtaposition*, *fusion*, *basicWord* o *isWord*.

Como se puede observar, SUG es una extensión de las DCG, por ello heredará muchas de sus características. La principal diferencia es que las reglas de producción son de la forma α++>β (en las DCG es $\alpha \rightarrow \beta$), y cada subconstituyente de β puede omitirse en la oración si se escribe entre el operador **<< >>**.

La gramática SUG que vamos a emplear en este trabajo es parcial en el sentido de que partiendo de un símbolo inicial únicamente se busca la derivación de sintagmas nominales, preposicionales y chunks verbales obviando cualquier otro tipo de constituyente. Esta misma gramática modificada y ampliada serviría para realizar análisis completos.

2.2. Sistema SUPP.

En la figura 1 se muestra el esquema general de procesamiento lingüístico que sigue el sistema SUPP.

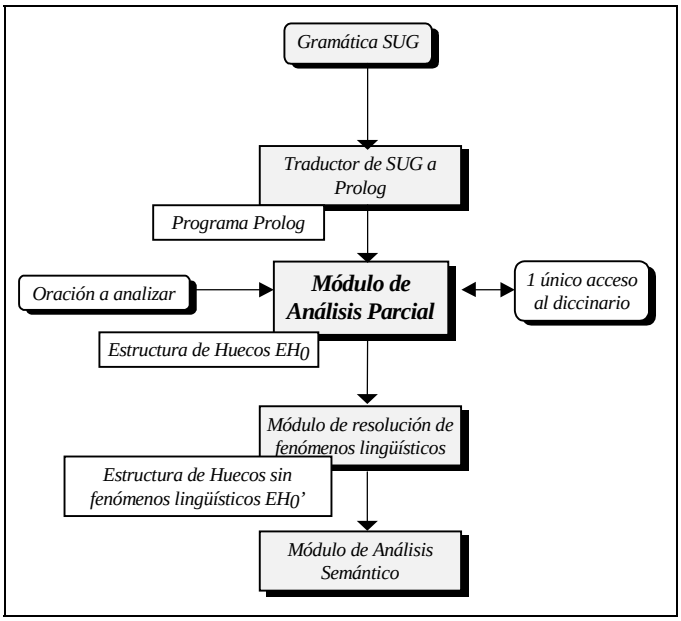

Figura 1. Sistema completo SUPP.

En primer lugar, se define una gramática SUG capaz de reconocer sintagmas nominales, sintagmas preposicionales y chunks verbales. Esta gramática se traduce automáticamente a cláusulas Prolog, obteniendo de esta forma el programa para el analizador. A partir de nuestro módulo de análisis sintáctico (parcial debido a la gramática introducida) se realiza un análisis sintáctico de la oración de entrada, empleando una técnica descendente, y obteniendo así su EH_0. Esta EH_0 almacenará únicamente los constituyentes relevantes encontrados en la oración de entrada y posteriormente se utilizará en el módulo de resolución de problemas lingüísticos. La solución consistirá en una nueva estructura de huecos, EH_0' en la que se han eliminado los problemas lingüísticos. Esta salida se puede emplear como entrada en distintas aplicaciones entre las que destacamos: obtención de las fórmulas lógicas mediante análisis semánticos, ligadura de subárboles sintácticos para la resolución de análisis sintácticos completos y aquéllas relacionadas con el campo de la recuperación de información y extracción de información.

2.3. Evaluación.

Para hacer las pruebas del sistema, aplicaremos nuestro módulo de análisis parcial sobre la salida de un etiquetador de manera similar al algoritmo propuesto en [Ken96]. Trabajaremos sobre el corpus etiquetado utilizado dentro del proyecto LESESP que consta, en su mayoría, de textos periodísticos y artículos de opinión sobre temas diversos: deportivos, políticos, humanos, etc., así como otro grupo de textos formados por breves narraciones literarias. Este corpus en castellano consta aproximadamente de 5M palabras etiquetadas con sus categorías gramaticales e información morfológica.

Tras aplicar nuestro analizador parcial SUPP sobre un fragmento del corpus formado por 71.849 palabras, dividido en 2738 oraciones y que ha sido corregido manualmente, hemos obtenido los siguientes resultados: para los sintagmas nominales simples (formados por un único núcleo de sintagma y constituyentes que lo complementan), se obtiene un 95% de precisión y un 94% de cobertura, mientras que para los sintagmas nominales completos o de mayor orden (no forman parte de ningún sintagma nominal de orden superior) el resultado es de un 80% de precisión y 79% de cobertura[2]. El descenso obtenido de los completos respecto a los simples es debido fundamentalmente a errores producidos por la acción de fenómenos lingüísticos, tales como la ambigüedad estructural, la coordinación, la elipsis, etc., irresolubles en el análisis parcial por la falta de información semántica en el corpus tratado y que deberán ser tratados en la fase posterior. Además, se detecta, que a pesar de tratarse de un corpus de texto corregido manualmente, existen algunos errores aislados de etiquetado léxico y morfológico que alteran el resultado tanto en el caso de los sintagmas nominales simples como en el de los completos, pudiendo, por tanto, obtenerse mejores resultados en corpus libres de errores de etiquetado.

3. Sistema EXIT.

EXIT es un sistema de extracción automática de información relevante a partir de textos de escrituras notariales de compraventa de inmuebles, así como su almacenamiento en una base de datos. La información relevante en este dominio son los datos de las personas que intervienen en la transmisión de los inmuebles (comprador/es, vendedor/es y notario), así como las características y situación del propio inmueble.

El sistema de extracción de información EXIT se desarrolló, inicialmente, utilizando lenguajes regulares para el análisis sintáctico y reconocimiento de entidades. Posteriormente, ha sido desarrollado utilizando reglas SUG en el reconocimiento de entidades debido a que esta técnica proporciona una serie de ventajas, como es la incorporación de información morfológica, sintáctica y semántica, gracias a su estructura de huecos, tal y como hemos mencionado anteriormente, y a la disminución del número de reglas debido a que las reglas SUG permiten la opcionalidad de algunos componentes, con lo que se mejora la eficiencia computacional del sistema.

```
<PLANTILLA-0001>: =
    DOC_NR          :
    CONTENIDO       :
<COMPRA_VENTA-0001>: =
    NOTARIO         :
    VENDEDOR        :
    COMPRADOR       :
    OBJETO_COMP     :
<PERSONA-0001>: =
    NOMBRE          :
    DNI             :
    DIRECCION       :
    TITULO          :
```

```
<ORGANIZACION-0001> : =
    NOMBRE          :
    CIF             :
    DIRECCION       :
    TITULO          :
<INMUEBLE-0001> : =
    IDENTIFICACION  :
    TIPO            :
    VIA             :
    DIRECCION       :
    LINDA           :
    REGISTRO        :
```

Figura 2. Plantillas del sistema EXIT**.**

Para la elaboración del sistema EXIT se han seguido las directrices marcadas en el MUC-7[3]. Así, en EXIT se distinguen: las tareas de reconocimiento de entidades, la resolución de correferencias, el

[2] Definimos la precisión como el cociente entre el número de sintagmas nominales analizados correctamente y el número de sintagmas nominales analizados en el texto. Consideraremos cobertura como el cociente entre el número de sintagmas nominales analizados correctamente y el número de sintagmas nominales reales en el texto.

[3] *MUC (Message Understanding Conference)*. Conferencia semestral cuyo objetivo es la evaluación de los sistemas de extracción desarrollados.

relleno de plantillas de elementos y la relación de plantillas. En la figura 2 se muestra un ejemplo de las plantillas que se han definido para ser rellenadas con la información relevante.

Para llevar a cabo las tareas anteriormente enunciadas se define, para el sistema EXIT, la arquitectura que se muestra en la figura 3, en la cual podemos distinguir las siguientes fases o etapas:

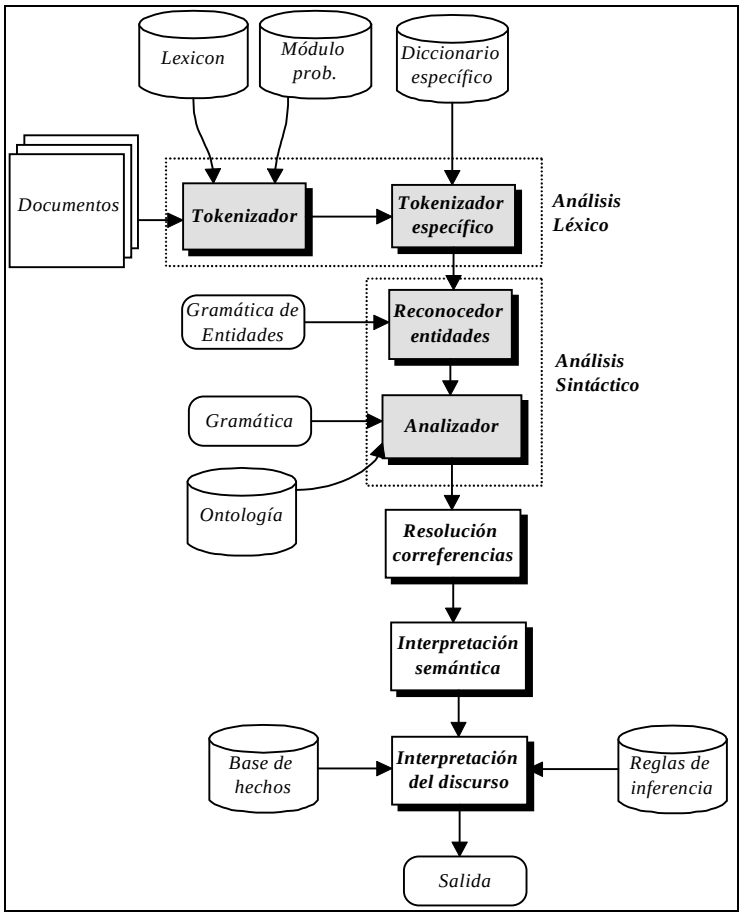

Figura 3. Arquitectura del sistema EXIT.

1. *Nivel de tokenización.* La información de entrada al sistema será una serie de escrituras de compraventa que se encuentran en formato electrónico. En primer lugar, se aplica un mecanismo para la determinación del límite de la frase [Muñ98]**.** Posteriormente, el tokenizador procesa cada una de las frases de la escritura, identificando palabras y asignando a cada una de ellas un identificador que permanece durante todo el proceso. Paralelamente, se procesa cada palabra, asignándole las categorías gramaticales posibles que pueda tomar dicha palabra. En esta etapa se utiliza principalmente un diccionario de propósito general; además de este diccionario usamos unos diccionarios específicos como son:
    - Diccionario de nombres (de 4337 entradas)
    - Diccionario de apellidos (de 4657 entradas)
    - Diccionario de localidades (de 53000 entradas)
    - Diccionario de actividades (de 1500 entradas)
2. *Análisis sintáctico*. En esta fase, se realiza un análisis sintáctico de las frases apoyándose en una gramática y en una ontología de rasgos, obteniéndose una serie de estructuras sintácticas. En esta etapa o módulo también se usa una gramática específica para identificar entidades o partes de entidades como es la dirección.
3. *Resolución de fenómenos lingüísticos*. A partir de las estructuras generadas en la etapa anterior se resuelven las correferencias existentes en el texto y se generan unas estructuras que no tengan correferencias. Para solucionar estos problemas se utilizan los métodos descritos en [Fer98].

4. *Interpretación del discurso*. Por último, se realiza una interpretación del discurso basándose en un modelo del mundo que se obtiene de una base de hechos y una serie de reglas de inferencia.

A pesar de las ventajas que incorpora el uso de reglas SUG, en el sistema EXIT hay una serie de carencias que no quedan eficientemente resueltas, como son, en primer lugar, que el sistema EXIT necesita una gramática de propósito general junto con la gramática de entidades para realizar el análisis sintáctico. Sin embargo, no es posible encontrar una gramática que trate todos los casos del lenguaje para obtener así una estructura sintáctica completa. Por otra parte, no existe ningún mecanismo en el sistema que permita identificar los sintagmas nominales definidos (sintagmas que hacen referencia a una entidad o a otro sintagma que ya ha aparecido en el texto) de todo el conjunto de sintagmas nominales identificados. Como consecuencia, no será posible resolver los casos de correferencia que evitarían la existencia de múltiples plantillas para una misma entidad.

4. Aplicación de SUPP a EXIT.

El objetivo de este trabajo se centra en la aplicación del analizador parcial SUPP descrito anteriormente al sistema de extracción de entidades definido en el sistema EXIT sobre el dominio restringido de las escrituras notariales de compraventa, solucionando las carencias que se han mostrado anteriormente.

La integración de SUPP en EXIT le proporcionará un único módulo capaz de analizar parcialmente la información relevante y reconocer las entidades. Por otra parte, este sistema será capaz de identificar los sintagmas nominales definidos, acrónimos y alias para aplicar posteriormente mecanismos de resolución de expresiones anafóricas. Con esto, conseguiremos tener una única plantilla para todos los sintagmas nominales que hacen referencia a una misma entidad.

Veamos un ejemplo con un fragmento de un texto notarial:

"...D. Juan López Pérez compra a la empresa MINSA S.A. el edificio sito en esta ciudad de Alicante y su calle Foguerer, número tres. El Sr. López se compromete a efectuar el pago mediante talón bancario a nombre del vendedor ...".

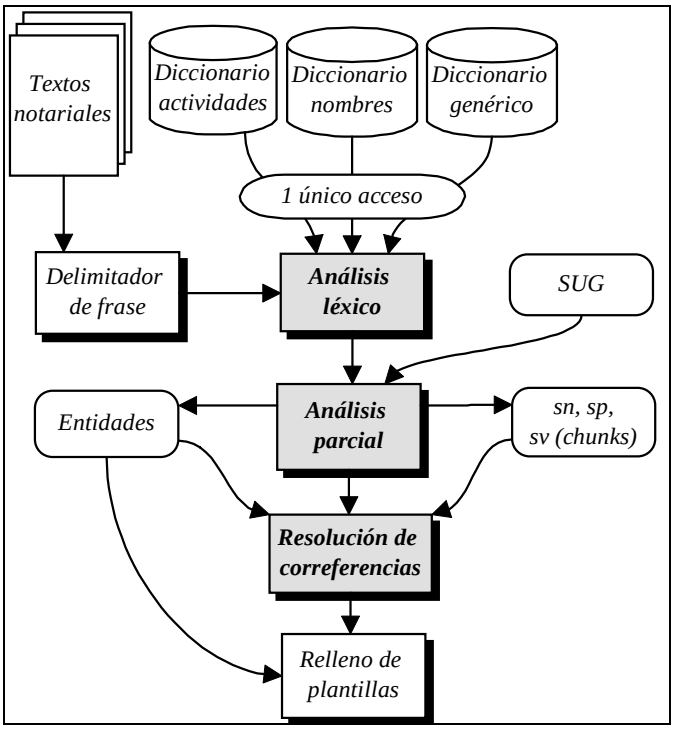

Figura 4. Aplicación de SUPP a Exit.

En este fragmento, nuestro sistema puede identificar 3 entidades: el comprador (D. Juan López Pérez), el vendedor (MINSA S.A.) y el inmueble, e identificar y resolver las expresiones anafóricas "el edificio", "el Sr. López", "el vendedor", "el pago".

Por tanto, el sistema propuesto es el que aparece en la figura 4, donde partiremos del conjunto de textos notariales de prueba que será preprocesado por el módulo delimitador de frase, obteniendo

oraciones separadas para su posterior procesamiento individual. Estas oraciones pasarán a través de un analizador léxico que mediante un único acceso a los diccionarios será capaz de etiquetar las palabras del texto con todas las categorías gramaticales encontradas para ésta, así como su información morfológica marcando además aquéllas que han sido reconocidas como nombres propios o actividades al consultar sus respectivos diccionarios.

El análisis parcial adquiere directamente la salida generada por el analizador léxico, y haciendo uso de la gramática SUG definida ad-hoc analiza todas las estructuras sintácticas correspondientes a sintagmas nominales, sintagmas preposicionales y chunks verbales que encuentre así como las entidades relevantes para el sistema EXIT (conjunto de nombres propios que forman un nombre de organización, persona o dirección). Como salida se obtiene un conjunto de entidades reconocidas y un conjunto con las estructuras sintácticas relevantes encontradas, entre las cuales se podrá percibir la existencia de algunos sintagmas nominales definidos que harán referencia a entidades reconocidas anteriormente (expresiones anafóricas).

Para resolver el problema generado por la correferencia se plantea el siguiente módulo del sistema, que tomando los sintagmas nominales afectados, será capaz de identificar sus referentes dentro del conjunto de entidades, tras lo cual, el sistema habrá adquirido toda la información que necesita para el relleno de las plantillas.

5. Conclusiones.

Se ha integrado el método de análisis parcial SUPP en el sistema EXIT resolviendo las carencias que planteaba este sistema, es decir, la necesidad de una gramática de propósito general que obtenga información relevante para el dominio, así como los problemas que planteaba el reconocimiento de sintagmas nominales definidos, acrónimos y alias, y la resolución de sus correferencias en el texto.

Lo que se pretende realizar con ello es la aplicación de estas técnicas a textos no restringidos, ya que consideramos que el sistema de análisis parcial está bien definido, como muestra su evaluación en textos de este tipo [Mar98b].

Referencias.

[Abn97] Abney, S. Part-of-Speech Tagging and Partial Parsing. In *Corpus-based Methods in Language and Speech Processing*. S. Young and G. Bloothooft, Eds., Kluwer Academic publishers. The Netherlands. 1997. pp. 119-136.

[Fer97] Ferrández, A.; Palomar, M.; Moreno, L. Slot Unification Grammar. In *Proceedings. of APPIA-GULP-PRODE* (Grado, Italy, 1997). pp. 523-532.

[Fer98] Ferrández, A., Palomar, M., and Moreno, L. Anaphor resolution in unrestricted texts with partial parsing. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, ACL'98 - COLING'98* (Montreal, Canada, 1998). pp. 385-391.

[Ken96] Kennedy, C; Boguraev, B. Anaphora for Everyone: Pronominal Anaphora Resolution without a Parser. In *Proceedings of the 16th International Conference on Computational Linguistics, COLING'96* (Copenhagen, Denmark, 1996). pp. 113-118.

[Llop98] Llopis, F.; Muñoz, R.; Suárez, A.; Montoyo, A. EXIT: Propuesta de un sistema de extracción de información de textos notariales. *Novatica*, *133* (Mayo-Junio 1998). pp. 26-30.

[Mar98a] Martínez-Barco, P.; Peral, J.; Ferrández, A.; Moreno, L.; Palomar, M. Analizador Parcial SUPP. In *Proceedings of VI biennial Iberoamerican Conference on Artificial Intelligence*, *IBERAMIA'98* (Lisboa, Portugal, 1998).

[Mar98b] Martínez-Barco, P.; Peral, J.; Navarro, B.; Ferrández, A.; Moreno, L.; Palomar, M. A Partial Parsing Strategy. In *Proceedings of Logical Aspects of Computational Linguistics, LACL'98* (Grenoble, France, 1998). Enviado, pendiente de aceptación.

[Muñ98] Muñoz, R.; Palomar, M. Sentence Boundary and Named Entity Recognition in EXIT system: Information Extraction System of Notarial Texts. In *Proceedings. of IV Int. Conference on Artificial Intelligence and Emerging Technologies in Accounting, Finance and Tax* (Huelva, Spain, 1998).

[Per80] Pereira, F.; Warren, D. Definite Clause Grammars for Language Analysis - A Survey of the Formalism and a Comparison with Augmented Transition Networks. *Artificial Intelligence*, *13* (1980).