Procesamiento del lenguaje Natural, nº 41 (2008), pp. 37-44 recibido 7-05-2008; aceptado 16-06-2008

Análisis sintáctico profundo del español: un ejemplo del procesamiento de secuencias idiomáticas*

Spanish deep parsing: the example of idiomatic sequences processing

Jorge Antonio Leoni de León, Sandra Schwab y **Éric Wehrli**
LATL - Departamento de Lingüística
Universidad de Ginebra
2, rue de Candolle
CH-1211 Ginebra 4,
Suiza
[jorge.leonideleon,sandra.schwab,eric.wehrli]@lettres.unige.ch

Resumen: En el Laboratorio de Análisis y de Tecnología del Lenguaje de la Universidad de Ginebra (Suiza), se ha desarrollado el analizador sintáctico profundo multilingüe FIPS, el cual es todavía un trabajo en progreso. Dicho analizador, inspirado de las teorías generativistas chomskyanas, se basa en la idea de conjuntos de estructuras sintácticas comunes a varios idiomas (ya sea a todas las lenguas o familias de lenguas). En este artículo presentamos una introducción a la estrategia general de FIPS, ejemplificada con el español, así como una muestra de aplicación al procesamiento de secuencias idiomáticas. Este tipo de secuencias, aunque generalmente procesadas como secuencias léxicas estáticas, pueden ser objeto de diversas transformaciones léxico-sintácticas, como la pronominalización clítica de un argumento interno o la substitución de elementos. Capturar el sentido de tales secuencias en la oración requiere una representación sintáctica profunda que permita establecer los vínculos entre la forma base y la realización (o forma superficial).
Palabras clave: analizador sintáctico profundo, expresiones idiomáticas

Abstract: FIPS, a multilingual deep parser, has been developed at the Language Technology Laboratory (LATL) of the University of Geneva (Switzerland).This parser, inspired by Chomskyan generative theories, is based on the idea that sets of syntactic structures are common to different languages (to all languages or to some language families). In this paper, we present an introduction to FIPS processing that we illustrate with Spanish and examples of multiword expressions. Such expressions, although generally processed as static lexical sequences, can indeed undergo various lexical-syntactic transformations, such as pronominalizations or substitutions. Retrieving such sequences' meaning requires a deep syntactic representation, which needs to establish the links between deep structures and surface forms.
Keywords: deep parsing, multiword expressions

1. Introducción

Desde hace varios años, en el *Laboratorio de Análisis y de Tecnología del Lenguaje* (LATL, 2008; Laenzlinger y Wehrli, 1991) de la Universidad de Ginebra se desarrolla el analizador sintáctico profundo multilingüe FIPS (Wehrli, 2004; Wehrli, 2007).[1] Este se inspira, fundamentalmente, del esquema teórico chomskyano (Chomsky, 1995, capítulo 1 con Howard Lasnik), con adaptaciones libres del modelo *Minimalista* (Chomsky, 2004), de *Simpler Syntax* (Culicover y Jackendoff, 2005) y de la *Gramática léxico-funcional* (Bresnan, 2001). Así, FIPS posee un núcleo gramatical común a todas las lenguas del sistema, al que se le agregan módulos especializados correspondientes a grupos de lenguas que presentan similitudes en cuanto a ciertos fenómenos, como por ejemplo los pronombres clíticos en las lenguas latinas. Esta estrategia reduce el tiempo necesario para la introducción de nuevas lenguas en el sistema, al haber un conjunto de condiciones y fenómenos sintácticos predefi-

* Esta investigación ha recibido el apoyo del *Fonds National Suisse pour la Recherche Scientifique* (*Fondo Nacional Suizo para la Investigación Científica*), proyecto $n^o 101412 - 103999$.

[1] Existe una versión en línea del analizador (LATL, 2008).

ISSN: 1135-5948

nidos, tanto para el total de lenguas, como para un subconjunto de ellas.

La ventaja de un analizador sintáctico profundo con respecto a los analizadores sintácticos superficiales, como Atserias et al. (2006), es su capacidad para identificar eficazmente las relaciones de distancia en la frase. Por ejemplo, los elementos constitutivos de las expresiones idiomáticas no siempre se encuentran próximos los unos de los otros, aunque está claro que la coocurrencia de dichos elementos es importante. Tal es el caso de la colocación "explotar un mito", la cual, aparte de su forma transitiva básica, puede encontrarse bajo una forma pasiva, "el mito ha sido explotado", o una forma nominal, "la explotación del mito". En este artículo describimos el funcionamiento de FIPS y abordamos de manera general sus ventajas en el procesamiento de expresiones idiomáticas.

2. *El analizador Fips*

La implementación de FIPS se ha concentrado en seis idiomas: alemán, español, francés, griego, inglés e italiano.[2] Sin embargo, otras lenguas también han sido tratadas, aunque parcialmente, como el rumano, el ruso, el polaco y el romanche sursilvano. La base de datos léxica de FIPS ha sido siempre una prioridad, de manera que el léxico ha ido alcanzado un notable nivel tanto cualitativo, como cuantitativo; el Cuadro 1 resume la cobertura léxica de FIPS en cifras absolutas:

Idioma	Lemas	Formas	Colocaciones
Inglés	54 000	90 000	5 000
Francés	37 000	227 000	12 500
Alemán	39 000	410 000	2 000
Italiano	31 000	220 000	2 500
Español	25 100	265 000	1500
Griego	12 000	90 000	225

Cuadro 1: Número de entradas en FIPSBD

De esta manera, la base de datos léxica de FIPS contiene lemas, que son las formas canónicas para acceder a las entradas léxicas, formas, que son todas las instancias declinadas o conjugadas de una entrada léxica, y colocaciones que se abordan en la sección 3.

Los análisis de FIPS requieren la conjunción de los resultados de tres sistemas interdependientes: una base de datos léxicos (FIPSBD), un etiquetador morfosintáctico (FIPSTG) y un analizador sintáctico (FIPSSYN).

Inspirándose de la gramática chomskyana, FIPS maximiza los rasgos gramaticales comunes a las lenguas a través de varios módulos que van del más general al más específico, siendo este último un conjunto de reglas propias a una lengua en particular (Wehrli, 2004).[3] Por ejemplo, el tratamiento de los pronombres clíticos ("le di el libro") de las lenguas latinas más el griego es procesado por medio del módulo *Romance* (Leoni de León y Michou, 2006).[4]

FIPSBD especifica, entre otros, los datos de subcategorización y selección, las funciones temáticas y los rasgos semánticos sintácticamente relevantes. Por ejemplo, en el caso de un verbo como "ver", tenemos la serie de valores parcialmente especificados en el Cuadro 2, donde "ID" se refiere al número único de identificación del verbo "ver" en la base de datos, "Inflexión" indica el paradigma de conjugación correspondiente y "Subcategorización" especifica las posiciones de sujeto y de objeto directo que deben estar ocupadas por un sintagma nominal ("NP"). Estas últimas posiciones están asociadas, respectivamente, con "Argumento 1" y "Argumento 2", donde las funciones gramaticales y temáticas son declaradas. Estas informaciones son provistas al utilizador por el etiquetador FIPSTG. Por otra parte, todas las formas posibles de una entrada léxica han sido introducidas en la base de datos por medio de un generador.

En el caso de las otras categorías gramaticales, la información guardada en FIPSBD puede ser muy similar; por ejemplo, ciertos adjetivos están subcategorizados ("orgulloso de ALGO"). Además, tenemos el caso de informaciones léxico-semánticas particularmente relevantes (rasgos de selección); por ejemplo, la propiedad $[+humano]$ es agregada a los sustantivos referidos a seres humanos a fin de dar cuenta del uso de

[2] En cuanto al procesamiento automático de la lengua española, podemos citar tanto el trabajo de La Serna (2004), como el de Bick (2008), éste último trata de un analizador basado en gramáticas de restricciones (*constraint grammar*).

[3] En el marco de la gramática sintagmática endocéntrica, HPSG, se han efectuado esfuerzos similares (LinGO Lab, 2008).

[4] El módulo *Romance* se encarga de: (i) la identificación de las secuencias clíticas; (ii) la asociación de dichas secuencias al verbo anfitrión (u otra categoría, según el idioma); (iii) la verificación de rasgos entre la secuencia clítica y los argumentos del verbo; y (iv) la interpretación de la secuencia clítica. La interpretación de las secuencias clíticas toma la forma de una categoría vacía en la posición de argumento, coindexada con el pronombre clítico en una posición más alta; la formación de una cadena entre ambos permite la corrobación de los rasgos pertinentes de caso y tema. El etiquetador de FIPS (FIPSTG) muestra los valores correspondientes (objeto directo, objeto indirecto, etc.) para cada vocablo de la oración.

Etiqueta	Valor
Lema	**ver**
ID	511005524
Inflexión	1
Subcategorización	`[NP_NP]`
Argumentos	
Argumento 1	
Función gramatical	`sujeto`
Argumento 2	
Función gramatical	`objeto directo`
Función temática	`tema`

Cuadro 2: FIPSBD: "ver"

la preposición española "a" para señalar un objeto directo referido a un humano. Así, (1) contrasta con (2), puesto que si bien tanto "estudiante" como "edificio" son objetos directos, en este último la preposición "a" está ausente:

(1) Vi al estudiante

(2) Vi el edificio.

Los valores parciales de "estudiante" en FIPSBD están dados en el Cuadro 3, dentro de los que se cuenta "`humano`. El valor "`noArg`" en "Subcategorización" indica que "estudiante" no tiene ningún elemento subcategorizado. En cambio, la entrada léxica de "edificio" (Cuadro 4) carece del rasgo "`humano`", pero posee el rasgo facultativo "`Objeto físico`". Según las informaciones de los Cuadros 3 y 4, FIPSSYN, que veremos más adelante, atribuye un valor de complemento directo a "al estudiante" luego del análisis de la preposición en función de la estructura del verbo y del sintagma nominal en cuestión.[5]

Etiqueta	Valor
Lema	**estudiante**
ID	511002166
Género	`masculino, femenino`
Número	`singular`
Inflexión	7
Rasgos	`humano`
Subcategorización	`noArg`

Cuadro 3: FIPSBD: "estudiante"

Ahora bien, si tenemos una oración como "vi el edificio", según las informaciones especificadas en el Cuadro 2, FIPSSYN va a intentar poner en relación el verbo "ver" con el sintagma nominal "el edificio", puesto que, según la subcategorización del verbo, la posición postverbal corresponde al objeto directo y éste debe ser un sintagma nominal. De esta forma, las informaciones para la combinación de los sintagmas son satisfechas.

[5]El tratamiento del objeto directo del español en FIPS, tanto para entes animados como inanimados, merece, sin duda alguna, ser abordado más en profundidad, en especial en lo que respecta a la fenomenología de la pronominalización clítica. Sin embargo, es un tema que exige más espacio del que podemos dedicarle aquí.

Etiqueta	Valor
Lema	**edificio**
ID	5110002787
Género	`masculino`
Número	`singular`
Inflexión	1
Rasgos	`Objeto físico`
Subcategorización	`noArg`

Cuadro 4: FIPSBD: "edificio"

FIPSSYN presupone la constitución de sintagmas endocéntricos consistentes en tres elementos: el núcleo del sintagma (`X`), a su izquierda una lista de subconstituyentes (`Izq`) y a su derecha otra lista de subconstituyentes (`Der`). Esquemáticamente lo representamos así:

[`Izq X Der`]

Cualquiera de estos elementos puede estar vacío. La variable "`X`" puede corresponder a cualquier categoría léxica: adverbio (`Adv`), adjetivo (`A`), complementador (`C`), determinante (`D`), interjección (`Inter`), preposición (`P`), sustantivo (`N`), Verbo (`V`). Además tenemos la categoría funcional de tiempo (`T`), que contiene toda la oración, así como una proyección funcional (`F`), usada para representar objetos predicativos, cuyo núcleo está constituido por un adjetivo, un adverbio, un sustantivo o una preposición. De esta manera una representación gráfica de un sintagma, o incluso de una oración, es necesariamente trinaria (Figura 1).

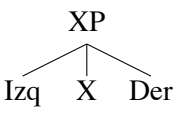

Figura 1: Estructura básica de FIPSSYN

Existen varios formatos de salida para los resultados de FIPS (texto, XML y con corchetes); todos consisten en una versión enriquecida de la frase original con delimitadores de los sintagmas, cuyo núcleo es un bigrama que denota la categoría gramatical a la que pertenecen ("`NP`" si es sustantivo, "`AP`" si es adjetivo, etc.). Para esta pre-

sentación, nosotros empleamos el formato basado únicamente en los corchetes (etiquetas sintagmáticas). Tomando en cuenta lo anterior, si introducimos en el sistema la frase "vi el edificio", obtenemos como resultado la versión etiquetada del ejemplo (3):

(3) [TP[DP] vi [VP [DP el [NP edificio]]]]

La Figura 2 representa gráficamente esta misma estructura:

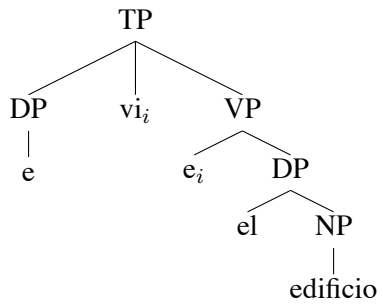

Figura 2: Representación de un análisis

Para facilitar la comprensión del análisis en (3), hemos agregado una categoría vacía, eventualmente coindexada con el verbo (vi$_i$ . . . e$_i$). Por su parte, FIPSTG resume las informaciones léxicas, presentadas en el Cuadro 5.

Vocablo	*vi*
Rasgos	VER-IND-PRS-1-SIN
ID único	511005524
Lema	**ver**
Vocablo	*el*
Rasgos	DET-SIN-MAS
ID único	511007887
Lema	**el**
Función	OBJ
Vocablo	*edificio*
Rasgos	NOM-SIN-MAS
ID único	511000278
Lema	**edificio**

Cuadro 5: Resultados del etiquetador

Las operaciones que le permiten a FIPS alcanzar estos resultados se apoyan en tres métodos: *Project*, *Merge* y *Move*.

2.1. Método *Project*

El método "Project" (*proyección*) crea un constituyente sintáctico sobre la base de un objeto léxico o de otro constituyente sintáctico. Todo elemento léxico identificado por FIPS a partir de las informaciones de FIPSBD es proyectado como un sintagma con un ítem léxico como núcleo.

En español, como se ve en el ejemplo (4), el artículo definido "el" proyecta un sintagma determinante (4a.), mientras que "edificio" proyecta un sintagma nominal (4b.). Los pronombres personales (4c.), que en FIPSBD son considerados como una forma especial de sustantivos, realizan lo que se llama una "metaproyección", es decir, proyectan inmediatamente su estructura superior, que en este caso es un DP (en FIPS todo sintagma nominal está contenido en un sintagma determinante). La metaproyección también es utilizada, en el análisis de las lenguas romances, para los verbos conjugados, que pasan a ser TP (4d.) (esta operación tiene como objetivo verificar la concordancia entre el sujeto y el verbo en las lenguas con sujeto desinencial).

(4) a. Determinantes:
el → [DP el]
b. Sustantivos:
edificio → [NP edificio]
c. Pronombres:
tú → [DP [NP tú]]
d. Verbos:
vi → [TP vi$_i$[VP e$_i$]]

En inglés (5a.), la metaproyección no tiene lugar, puesto que esta operación no se ve justificada dada la pobreza morfológica de ese idioma. También hay lenguas que requieren una metaproyección más compleja, como en el alemán (5b.) que en nuestro esquema necesita una metaproyección superior al sintagma de tiempo (TP) para dar cuenta del fenómeno del verbo en posición final de oración, que es considerada como su posición canónica.

(5) a. Inglés:
reads → [TP [VP reads$_i$]]
b. Alemán:
regnet[6] → [CP regnet$_i$[TP [VP e$_i$]]

2.2. Método *Merge*

El método "Merge" es el mecanismo de combinación sintagmática de FIPS. Cada vez que el analizador lee un vocablo, éste es transformado en un constituyente, es decir, en una proyección como las explicadas en la sección 2.1. La proyección puede ser combinada ("merged") con constituyentes completos o parciales en cualquiera de sus contextos. En ese momento, se abren dos posibilidades: una agregación a la izquierda o una agregación a la derecha.

[6] En español, "llueve".

Una agregación a la izquierda es el caso típico del sujeto y el verbo. Esta consiste en la inserción de un constituyente en el contexto izquierdo de otra proyección, con la que es compatible. Por ejemplo, en (6), el pronombre personal (6a.), luego del reconocimiento del verbo (6b.), es agregado como un subconstituyente izquierdo de la nueva proyección verbal (es decir, como un sujeto), obteniendo así (6c.).

(6) a. ella → [$_{DP}$ ella]
b. duerme → [$_{TP}$ duerme [$_{VP}$]]
c. [$_{TP}$ [$_{DP}$ ella] duerme [$_{VP}$]]

Por el contrario, una agregación a la derecha corresponde a la situación en la que una proyección es agregada como un subconstituyente derecho de su propio contexto izquierdo. Este es el caso típico de los sintagmas determinantes, en los que el sintagma nominal es insertado a la derecha del sintagma determinante (DP); dicho de otra forma, los sintagmas determinantes acogen un sintagma nominal a la derecha del núcleo del constituyente:

Por ejemplo, en (7), el vocablo "el" proyecta un constituyente DP (7b.), en la gramática de FIPS, los DP ocupan una posición superior a los NP. En otras palabras, un DP puede puede tener un NP como argumento. De esta manera, la proyección (7c.) es combinada con (7b.) (es decir, introducida a la derecha de esta última), lo que produce el sintagma (7d.). El procedimiento para satisfacer los argumentos de un verbo son básicamente los mismos. Así, una vez reconocido el sintagma (7a.), el DP se incorpora a la derecha del sintagma verbal. El resultado de toda esta operación lo tenemos en (7e.).

(7) a. vi → [$_{TP}$ vi [$_{VP}$]]
b. el → [$_{DP}$ el]
c. edificio → [$_{NP}$ edificio]
d. [$_{DP}$ el [$_{NP}$ edificio]]
e. [$_{TP}$ [$_{DP}$] vi [$_{VP}$ [$_{DP}$ el [$_{NP}$ edificio]]]]

La operación "Merge" debe ser validada ya sea según las propiedades léxicas, como los rasgos de selección, o según ciertas propiedades generales (como por ejemplo los adverbios, las adjunciones y los paréntesis que pueden modificar libremente las proyecciones).

Según el Cuadro 2, el verbo "ver" se combina con un sustantivo en posición postverbal, que es un objeto directo, mientras que en posición preverbal, se combina con otro sustantivo, que es un sujeto con el que debe verificar los rasgos de persona y número (aunque en lenguas como el español, dicha posición puede estar vacía). Dado que en posición postverbal tenemos el sintagma "el edificio", reconocido como un sintagma nominal (por lo tanto compatible con las informaciones de "ver"), FIPS lo reconoce en esta posición como un objeto directo.

2.3. Método *Move*

La estructura general de superficie es el resultado de la combinación de las operaciones de "Project" y "Merge". Sin embargo, es necesario un mecanismo adicional para satisfacer las condiciones de uniformidad como, por ejemplo, la asignación de funciones temáticas. Tal es el objetivo del método "Move" ("mover"), el cual maneja la relación de elementos extraídos o dislocados con las posiciones que ocupaban originalmente. Un caso típico es el de las oraciones interrogativas parciales, como la oración inglesa en (8):

(8) a. Who did you invite ?
b. [$_{CP}$[$_{DP}$ who]$_j$ did$_k$ [$_{TP}$[$_{DP}$ you] e$_k$ [$_{VP}$ invite e$_j$]]]

El método "Move" consiste en la creación de una cadena de coindexaciones. En el ejemplo (8b.) tenemos dos elementos desplazados: el pronombre interrogativo "who" y el auxiliar "did". Dos hechos justifican la utilización de este mecanismo para el pronombre "who". En primer lugar el pronombre "who", para ser interpretado correctamente, necesita estar asociado a un verbo, el cual se encuentra lejos en la frase; por este motivo, su interpretación es diferida y el pronombre es insertado en una estructura temporal (en una pila). Luego, el verbo necesita satisfacer tanto su subcategorización (_NP), como la asignación de caso y función temática correspondiente. Aunque la posición postverbal se encuentra vacía, en la pila tenemos un elemento que cumple los requisitos para ser interpretado con respecto al verbo. Entonces una cadena de categorías vacías ("e") coindexadas es creada entre la posición de argumento (postverbal) y el pronombre "who". En segundo lugar, tenemos la creación de una correferencia entre el auxiliar "did" y su posición de origen. En este caso se trata de la manera de representar la inversión del sujeto en las interrogativas, fenómeno típico del inglés.

2.4. Ejemplo completo

Consideremos el análisis de la oración "Ana vio el edificio" a fin de ilustrar los mecanismos descritos:

Etapa 1 El analizador lee "Ana" y *meta-proyecta* la estructura [$_{DP}$[$_{NP}$Ana]].

Etapa 2 El analizador lee "vio" y metaproyecta una estructura de frase $[_{TP}\ vio_i\ [_{VP} e_i]]$.

Etapa 3 Una operación de "Merge" es efectuada entre el TP y el DP, que será ubicado a la izquierda de la proyección de tiempo: $[_{TP}\ [_{DP}[_{NP} Ana]]\ vio_i\ [_{VP}\ e_i]]$.

Etapa 4 El parser identifica el determinante "el" y *proyecta* la estructura $[_{DP} el]$.

Etapa 5 Una operación de "Merge" es efectuada entre el sintagma TP de la izquierda y el DP identificado; $[_{DP} el]$ es agregado a la derecha del TP: $[_{TP}\ [_{DP}[_{NP} Ana]]\ vio_i\ [_{VP}\ e_i\ [_{DP} el]]]$.

Etapa 6 El parser identifica el sustantivo "edificio" y *proyecta* la estructura $[_{NP} edificio]$.

Etapa 7 Una operación de "Merge" es efectuada entre el sintagma DP derecho del TP, en el que "edificio" es agregado como constituyente derecho del DP "el": $[_{TP}\ [_{DP}[_{NP} Ana]]\ vio_i\ [_{VP}\ e_i\ [_{DP}\ el\ [_{NP} edificio]]]]$.

La última etapa produce la estructura completa.

3. FIPS *y el reconocimiento de expresiones idiomáticas: una propuesta*

Dentro del marco de las tecnologías desarrolladas en el LATL (2008), se cuentan varias investigaciones sobre el procesamiento de las expresiones idiomáticas y de las colocaciones. Por ejemplo, Nerima, Seretan, y Wehrli (2006) y Seretan (2008) utilizan un procedimiento híbrido multilingüe, sintáctico-estadístico, para la extracción y el reconocimiento de las colocaciones. Por otra parte, Leoni de León (2008) ha trabajado en una propuesta de representación léxico-sintáctica orientada a reconocer y reproducir el funcionamiento de las expresiones idiomáticas, desde una perspectiva más próxima a la lexicografía. Estas propuestas abordan las interfaces entre el léxico y la sintaxis desde una perspectiva computacional. En la misma línea, es interesante citar también el sistema de asistencia terminológica TwicPen (Wehrli, 2006), que permite limitar el número de traducciones entre dos pares de lenguas sobre la base de un análisis lingüístico de un texto seleccionado para su traducción. TwicPen explota los recursos morfosintácticos de FIPS (así como las lenguas disponibles), aunados a un procesamiento sintáctico de las colocaciones, lo que permite recuperar estas unidades aún en circunstancias en que sus elementos constitutivos se encuentran morfológicamente modificados o mantienen relaciones de distancia. Todas estas investigaciones están en progreso, aunque ya dieron lugar a algunas publicaciones (ya mencionadas).

Las expresiones idiomáticas, a menudo consideradas como elementos estáticos, pueden presentar una morfosintaxis bastante rica (Leoni de León, 2008). Un buen ejemplo es la expresión idiomática "meter la pata", corriente en el español coloquial. Esta expresión se caracteriza por presentar casi todas las opciones sintácticas posibles para una expresión idiomática. Por ejemplo, el núcleo (verbal) de "meter la pata" puede ser nominalizado (9a.) o bien su argumento interno puede ser pronominalizado (9b.) en un contexto discursivo, operación que implica la adjunción de un complemento:

(9) a. Metida de pata.
b. La metió hasta el fondo.

Estas operaciones son difícilmente tomadas en cuenta en los sistemas de extracción estadísticos, impresión reforzada por las relaciones de concordancia entre el núcleo de una expresión adjetiva y un sustantivo. Por ejemplo, en la secuencia "hecho polvo" es el participio el que hace la concordancia de género y número, mientras que el colocativo no sufre modificación alguna:

(10) a. Él estaba hecho polvo.
b. Ella estaba hecha polvo.

No está de más agregar que la expresión "hecho polvo" proviene en realidad de la forma verbal "hacer polvo". Esto es una muestra de una relación transcategorial que va de una forma verbal a una forma adjetiva. De esta manera tenemos dos fenómenos idiomáticos que presentan relaciones de distancia ya sea entre sus elementos constitutivos, como en "meter la pata", o que no sólo pueden manifestarse con categorías diferentes (la forma verbal "hacer polvo" se convierte en un adjetivo, "hecho polvo"), sino que además pueden concordar en género y número, por ejemplo. La arquitectura de FIPS permite recuperar muchos de estos fenómenos.

En el caso de las pronominalizaciones clíticas, como el ejemplo (9b.), la identificación de la expresión como una instancia de "meter la pata" requiere el establecimiento de la relación entre el pronombre clítico de objeto directo, "la", y la posición de argumento, la cual estimamos ocupada

por una categoría vacía coindexada con el clítico. La adjunción de un complemento circunstancial ("hasta el fondo" en este caso) debe contar dentro de la base de conocimientos idiomáticos, como lo señala Leoni de León (2008). En lo que respecta a la expresión en (10), el punto fundamental está en la necesidad de establecer una relación entre el participio y el elemento nominal al cual se refiere, con independencia del sustantivo "polvo".

Las expresiones idiomáticas son relativamente fáciles de identificar, cuando su realización es lineal. Tal es el caso del ejemplo (11), para el que FIPS produce el análisis en (12). Sabemos que la expresión "romper un récord" ha sido correctamente identificada por FIPS, debido a que FIPSTG indica el valor "541001721" de la etiqueta "Colocación", que es el número de identificación único de esta expresión en FIPSBD (Cuadro 6). Por otra parte, FIPS tampoco tiene dificultades para identificar dicha expresión, incluso si el artículo indefinido "un" es sustituido por el artículo definido "el"; para esto ha bastado indicar en FIPSBD que la expresión necesita la presencia de un artículo.

(11) Él rompió un récord.

(12) [$_{TP}$[$_{DP}$Él] rompió [$_{VP}$[$_{DP}$ un [$_{NP}$ récord]]]]

Vocablo	*rompió*
ID	511005410
Lema	**romper**
Colocación	541001721
Vocablo	*un*
ID	511007415
Lema	**un**
Función	OBJ
Vocablo	*récord*
ID	511001433
Lema	**récord**
Colocación	541001721

Cuadro 6: Valores de una expresión transitiva

Ahora bien, la capacidad de FIPS para reconocer la expresión (11) no se ve alterada aunque el objeto directo esté modificado por un sintagma preposicional ("Él rompió el récord de Claudia") o, incluso, si, además, la expresión está realizada como una oración pasiva, "El récord de Claudia ha sido roto". De esta forma, como lo muestran tanto el análisis en (13), como los resultados de FIPSTG (Cuadro 7), FIPS no tiene ninguna dificultad para reconocer una expresión, aunque se hayan establecido relaciones de distancia. Esto se consigue, por un lado, con la creación de una cadena de coindexaciones que va de la categoría vacía en posición postverbal, "[$_{DP}$e$_i$]" hasta el sintagma determinante que contiene el sujeto "El récord de Claudia", por otro lado, el análisis profundo de FIPSSYN, identifica el sintagma preposicional "de Claudia", como un subconstituyente del sintagma determinante sujeto, "El récord".

(13) [$_{TP}$[$_{DP}$ El [$_{NP}$ récord [$_{PP}$ de [DP Claudia]]]]$_i$ ha [$_{VP}$ sido [$_{VP}$ roto [$_{DP}$e$_i$]]]]

Vocablo	*el*
ID	511007887
Lema	**el**
Función	SUBJ
Vocablo	*récord*
ID	511001433
Lema	**récord**
Colocación	−541001721
Vocablo	*roto*
ID	511005410
Lema	**romper**
Colocación	−541001721
Función	SUB:récord

Cuadro 7: Valores de una expresión pasiva

Dentro de los valores del Cuadro 7, encontramos "SUB:récord" para "roto". Este valor indica que el analizador reconoció el lema como sujeto de "romper"; además, este valor se encuentra también asociado a la forma pasiva del verbo, de manera que la información es fácilmente recuperable. Se trata de una información referida al sujeto gramatical de la oración.

Las posibilidades de FIPS para el tratamiento de las expresiones idiomáticas son inmensas, es así como existe otra estrategia, (Leoni de León, 2008) que consiste en la proposición de un formalismo correlacional, llamado *Tsool*, que codifica el comportamiento morfosintáctico de las expresiones idiomáticas. Dicho formalismo es reproducido computacionalmente en un sistema (llamado *Mulkin*) que interactúa con FIPS para explotar los análisis sintácticos de este sistema, a fin de poder conjugar los análisis con las informaciones fraseológicas almacenadas, y así reconocer las expresiones idiomáticas. Tanto *Tsool* como *Mulkin* se encuentran en una etapa temprana de desarrollo, y, como ya lo señalamos oportunamente, ambos se orientan hacia una representación más cercana de la lexicografía. Dentro de los elementos considerados podemos citar las relaciones de rima, las posibilidades de conmutación y de permutación de las expresiones. Una de las aplicaciones previstas para este sistema es

la filtración de secuencias luego de una operación de extracción a partir de corpus de gran tamaño.

4. Conclusión

FIPS es un analizador sintáctico capaz de identificar las relaciones profundas entre los constituyentes de la oración. La arquitectura multilingüe de FIPS, basada en una serie de módulos especializados en conjuntos de fenómenos sintácticos por familias o grupos de lenguas, facilita la inclusión de nuevas lenguas en el sistema, maximizando la utilización del código de la aplicación. Las propiedades de FIPS se muestran particularmente útiles en el reconocimiento de secuencias idiomáticas, puesto que estas no son necesariamente estáticas, sino que pueden ser objeto de modificaciones, por las cuales sus constituyentes no se realizan linealmente, sino de manera discontinua (relaciones de distancia).

Bibliografía

Atserias, Jordi, Bernardino Casas, Elisabet Comelles, Meritxell González, Lluís Padró, y Muntsa Padró. 2006. Freeling 1.3: Syntactic and semantic services in an open-source nlp library. En *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006), ELRA.*, Génova, Italia, Mayo.

Bick, Eckhard. 2008. A constraint grammar parser for spanish. Página web. [Dirección electrónica: http://beta.visl.sdu.dk/pdf/TIL2006.pdf ; Visitada el 2 de mayo de 2008].

Bresnan, J. 2001. *Lexical Functional Syntax*. Blackwell, Oxford.

Chomsky, Noam. 1995. *The Minimalist Program*. MIT Press, Cambridge.

Chomsky, Noam. 2004. Beyond Explanatory Adequacy. En A. Belletti, editor, *The Cartography of Syntactic Structures*. Oxford University Press, Oxford.

Culicover, Peter y Ray Jackendoff. 2005. *Simpler Syntax*. Oxford University Press, Oxford.

La Serna, Nora. 2004. Un analizador sintáctico eficiente para gramáticas del español. *Rev. investig. sist. inform.*, 1(1):19–26.

Laenzlinger, Christopher y Éric Wehrli. 1991. FIPS : Un analyseur interactif pour le français. *TA Informations*, 32(2):35–49.

LATL. 2008. Laboratoire d'Analyse et de Technologie du Langage. Página web. [Dirección electrónica : http://www.latl.unige.ch/ ; Visitada el: 28 de abril de 2008].

Leoni de León, Jorge Antonio. 2008. *Modèle d'analyse lexico-syntaxique des locutions espagnoles*. Tesis en lingüística, Université de Genève, Ginebra, Suiza, Mayo.

Leoni de León, Jorge Antonio y Athina Michou. 2006. Traitement des clitiques dans un environement multilingue. En Piet Mertens Cédrick Fairon Anne Dister, y Patrick Watrin, editores, *Verbum ex machina : Actes de la 13e conférence sur le traitement automatique des langues naturelles (TALN 2006)*, volumen 1 de *Cahiers du Cantal 2.1*, páginas 541–550, Louvain-la-Neuve, Belgique, 10-13 avril. Association pour le Traitement Automatique des Langues, UCL Presses Universitaires de Louvain.

LinGO Lab, CSLI. 2008. CSLI Linguistic Grammars Online. Página web. [URL: http://lingo.stanford.edu/ ; Visitada el 2 de mayo de 2008].

Nerima, Luka, Violeta Seretan, y Éric Wehrli. 2006. Le problème des collocations en TAL. *Nouveaux cahiers de linguistique française*, (27):95–115.

Seretan, Violeta. 2008. *Collocation Extraction in Syntactic Parsing*. Ph.D. tesis, Université de Genève, Juin.

Wehrli, Éric. 2004. Un modèle multilingue d'analyse syntaxique. En Antoine Auchlin Marcel burger Laurent Filliettaz Anne Grobet Jacques Moeschler Laurent Perrin, y Corinne Rossari et Louis de Saussure, editores, *Structures et discours : Melanges offerts à Eddy Roulet*, Langue et pratiques discursives. Éditions Nota bene, Canada, páginas 311–332.

Wehrli, Éric. 2006. Twicpen: hand-held scanner and translation software for non-native readers. En *Proceedings of the COLING/ACL on Interactive presentation sessions*, páginas 61–64, Morristown, NJ, USA. Association for Computational Linguistics.

Wehrli, Éric. 2007. Fips, a "Deep" Linguistic Multilingual Parser. En *ACL 2007 Workshop on Deep Linguistic Processing*, páginas 120–127, Prague, Czech Republic, Juin. Association for Computational Linguistics.