Káñina, Rev. Artes y Letras, Univ. Costa Rica. Vol. XXX (1), pág. 83-98, 2006
ISSN: 0378-0473

SISTEMA LÉXICO VIRTUAL

Jorge Antonio Leoni de León[*]

ABSTRACT

Departing from the concept of lexicon as a system and based on the Government and Binding Theory , the idiosyncratic pieces of information of nominal lexical structures aimed to be projected syntactically are identified. It is assumed the existence of a determining sintagma in which operations of quantification, of number attribution and, eventually, of genre take place. This is shaped in the eXtensible Markup Language (XML) to be exploited in systems of automatic treatment.
Key words: lexicography, Government and Binding, eXtensible Markup Language (XML).

RESUMEN

Partiendo del concepto de léxico como sistema, y siguiendo la Teoría de Rección y Ligamento, se identifican las informaciones idiosincráticas de las estructuras léxicas nominales, destinadas a proyectarse sintácticamente, se asume la existencia la existencia de un sintagma determinante en el que se realizan operaciones de cuantificación, de atribución de número y, eventualmente, de género. Esto se modeliza en el Lenguaje de etiquetado extensible para ser explotado en sistemas de tratamiento automático.
Palabras claves: lexicografìa, Rección y Ligamento, Lenguaje de etiquetado extensible.

1. Introducción

La lingüística es el estudio de un sistema sígnico complejo que permite a los seres humanos transmitir ideas, deseos y necesidades. Los elementos que conforman dicha complejidad están constituidos por otros sistemas que interactúan de diversas maneras; generalmente, se ha tendido a pensar que unos se encargan de almacenar información y otros, de elaborarla. Ahora bien, para comprender nuestra busca, es necesario recordar que, lingüísticamente, *sistema* se entiende como un conjunto ordenado de elementos relacionados entre sí, cuyo valor es construido diferencialmente; por otra parte, la *virtualidad* se concibe, también, como la cualidad de contener en sí todas las potencialidades necesarias para ejecutar una determinada acción, en nuestro caso, manifiesta por medio del habla. De esta forma, un sistema virtual es la consecuencia de la identificación y diferenciación de las cualidades de una entidad, de la que sólo percibimos sus fenónemos, y que en nuestro caso es el *léxico.*

Definir *léxico* no es tarea fácil. Ya el concepto de *palabra*, como unidad mínima del léxico, es de por sí una idea elusiva a toda exposición de sus características que aspire a ser clara y exacta. Incluso la representación del léxico como un conjunto ordenado de elementos (*palabras*) enfrenta serias dificultades relacionadas con la identidad de sus unidades mínimas, las cuales

* Profesor de la Escuela de Filología, Universidad de Costa Rica. Correo electrónico: kilimanjaro@racsa.co.cr

se clasifican en categorías (gramaticales) separadas, que cumplen funciones complementarias distintivas en la oración y que están compuestas por otras unidades menores. Una buena manera de emprender una definición de léxico es comenzar por el concepto de gramática como sistema formal constituido por un conjunto de reglas explícitas mecánicamente aplicables, que transforman una cadena de símbolos (de entrada) en otra cadena de símbolos (de salida).[1] En esta primera aproximación, nuestra gramática, basada en la *Teoría de Rección y Ligamento* y en los avances hechos dentro de lo que se conoce como el *Programa Minimalista*, plantea que la cadena de símbolos de entrada está constituida por el léxico y la cadena de símbolos de salida, por el vocabulario. El léxico, de esta forma constituido en componente básico de la gramática, provee la información semántica, sintáctica y fonológica necesaria para la correcta operación de las reglas de estructura de la frase.[2] Cada elemento léxico especifica su propia estructura morfo-fonológica, la cual, claro está, debe respetar los *parámetros* de la lengua a la cual pertenece, además de los *principios* de la *Gramática Universal*. En otros términos, la interpretación de los enunciados se lleva a cabo a partir del vocabulario a través de la interacción de las reglas de la gramática (sintaxis) y del léxico; mientras que la generación consiste en la satisfacción de las reglas de la gramática, según la secuencia de informaciones que contengan las unidades léxicas extraídas; es decir, que el léxico sería un inventario de ítemes básicos sobre los cuales operan las reglas de la gramática con restricciones sobre su libre aplicación. Tal y como lo hemos propuesto hasta este momento, estamos ante un modelo en el cual la relación del léxico y la gramática se asemeja mucho a la de la Ilustración 1.

En este modelo (Ilustración 1), las reglas de la gramática se aplican según la información provista por los ítemes léxicos. Sin embargo, estos, al contener información gramatical (sintáctica, semántica, fonológica), pertenecerían también al conjunto de mecanismos que forman la gramática. En otros términos, las reglas, por sí solas, no son la gramática, sino que ésta estaría constituida por las reglas y el léxico. Por lo tanto, cada parte

Relación autonómica del léxico y la sintaxis según una perspectiva tradicional

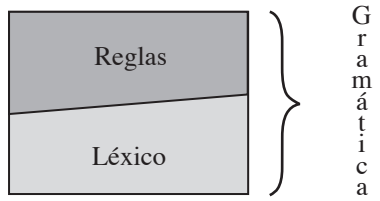

Ilustración 1

es autónoma y la interacción entre ellos se limita a una simple transferencia de información, esto en contradicción con la naturaleza de los datos codificados en el léxico y con la manera en que la sintaxis los utiliza, como se desprende del *Principio de Proyección*, citado a continuación:

1. Principio de Proyección

La información léxica debe estar representada a todos los niveles de la gramática.

Este principio establece un vínculo claro entre el léxico y la sintaxis. La información a la que alude es de naturaleza diversa, comúnmente se acepta que se refiere a los datos señalados en (2):

2. Información léxica idiosincrática

- Sentido / Significado: conocimiento sobre la realidad.
- Categoría sintáctica: adjetivo, sustantivo, verbo, preposición, etcétera.
- Rasgos gramaticales: número, persona, aspecto, entre otros.
- Clasificación morfológica: raíz, morfema.
- Morfología derivacional: asignación de morfemas compatibles.
- Subcategorización: información configuracional.
- Estructura de argumento y predicado: por ejemplo, verbos de marcaje especial (ECM), ante y posposición, roles temáticos (roles-θ).

- *Caso abstracto: argumentos*.
- *Registro: estilo*.

Está claro que la mayor parte de estas informaciones tienen implicaciones en la organización de sintagmas en la frase; y que, además, hay operaciones cuya ejecución dependen de la interrelación entre estos datos y los módulos de la sintaxis.

Siguiendo el principio citado en (1), y tomando en cuenta las propiedades y las dependencias que se desprenden de él, nuestra gramática se aproxima más a lo presentado en la Ilustración 2.

Relación intedependiente de la sintaxis y el léxico según el Principio de Proyección

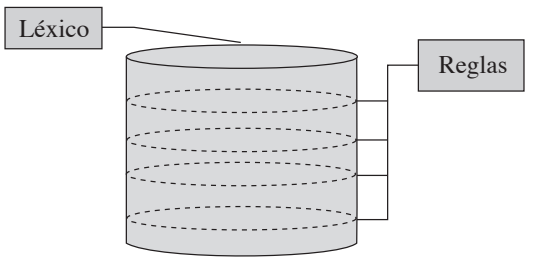

Ilustración 2

En esta ilustración se grafica la conexión planteada entre el léxico y las reglas de la sintaxis en el sentido de que las reglas, para realizarse, necesitan de la información léxica y esta información requiere de las reglas para la producción del enunciado; este proceso se da secuencialmente según la jerarquía de los módulos sintácticos y la pertinencia de los datos.

Recapitulando lo que hemos dicho hasta este momento, *sistema léxico virtual* alude, por tanto, a un conjunto ordenado de elementos relacionados entre sí, de valor calculado diferencialmente, el cual contiene todas las potencialidades para la manifestación fenoménica de las unidades que lo conforman y que constituyen subconjuntos de informaciones idiosincrásicas. La naturaleza de dichas relaciones está determinada por los requerimientos de las reglas de la gramática. Nuestro objetivo es desarrollar un sistema de codificación del léxico que permita obtener un nivel óptimo de desambiguación en la realización de todos los lemas coordinados realizados, a partir de las características que se puedan deducir de la sintaxis. Esto debe ser entendido como la transformación de unidades léxicas mediante un conjunto de reglas y marcas, por cuanto codificar implica transformar mediante las reglas de un código la formulación de un contenido, donde código es comprendido como un sistema de signos y de reglas que permite formular y comprender dicho sistema de transformación de unidades léxicas, el cual se espera que minimice los efectos de la ambigüedad en la oración. Por cuanto la ambigüedad, característica inherente al lenguaje humano, es imposible de reducir de manera determinista, nosotros aspiramos a reducir sus efectos en la interpretación de la proyección frástica de las cabezas léxicas.

Debido a la complejidad de fenómenos relacionados con el léxico, hemos debido limitarnos a una sola categoría como foco de nuestra investigación, en otra ocasión emprenderemos una exploración más completa de las redes de relaciones intercategoriales. De esta forma, intentaremos identificar estructuras sintácticas como resultado de la información contenida en el léxico; esto debería permitir, en cierta medida, proponer una técnica de desambiguación a partir de nuestro aparato teórico y con el fin de mejorar el tratamiento automático de estas formas, gracias a la detección de los mecanismos de proyección sintáctica de los sintagmas que analizamos. Desde nuestra perspectiva, claro está, la teoría lingüística es pertinente en el tratamiento automático del lenguaje natural.

Tomaremos como base de la información léxica sintácticamente relevante aquella que señalamos en (2), con excepción de aquellos ítemes que se ubican fuera de nuestro ámbito, como, por ejemplo, *registro* y *sentido/significado*.[3] En cuanto a la morfología, asumimos la postura lexicalista, que propone el almacenamiento de estructuras complejas en el léxico, que interactúan con un módulo de reglas morfológicas. En este sentido, siguiendo una propuesta de Lorenzo González (1995: 37-49, 112-118), ciertos lemas

pueden carecer de marcas de género, mientras que otros, pueden derivarlos sintácticamente. Los morfemas de número, en cambio, sí serían adquiridos por los lemas a lo largo del proceso de generación. Es necesario recordar que la elaboración morfológica está fuera del ámbito de esta investigación por cuanto nos limitamos a datos contenidos en el léxico.

La representación informática de estructuras del lenguaje natural presenta varias dificultades, entre las que podemos citar la forma de los datos mismos, su almacenamiento y su tratamiento. A menudo, estos tres aspectos responden a conceptos y técnicas muy alejadas una de la otra. Sin embargo, es posible alcanzar una armonía que permita tratar las informaciones a partir de su representación, lo cual encontramos en el conjunto de tecnologías asociadas al *Lenguaje de etiquetado extensible* (XML por sus siglas en inglés), de lo que trataremos en su debido momento.

2. Las estructuras nominales

Como ya lo hemos señalado, en esta investigación nos concentramos en las categorías nominales, en especial en su estructura de argumentos, base del procesamiento sintáctico. Es necesario indicar además que nosotros asumimos la *Hipótesis del DP*[4], también conocida como *Hipótesis de Abney* (Abney, 1987), la cual propone un paralelismo entre la estructura de las formas verbales y la de las formas nominales, de tal suerte que así como los sintagmas verbales (VP) son argumentos del sintagma de la inflexión (IP) (ver 3), los sustantivos son argumentos de una estructura superior que corresponde a los determinantes, ilustrado en (ver el ejemplo 5 y la Ilustración 4)

3. $[_{IP}\ [_{I},\ [_{I^\circ}\ \alpha]\ [_{VP}\ \beta]]]$
$[_{IP}\ [_{I},\ [_{I^\circ}\ \mathbf{ha}]\ [_{VP}\ \mathbf{salido}]]]$

La estructura en (3) simboliza la relación entre unidades de valor funcional y de valor léxico; es la conexión que existe entre un auxiliar (como *haber*) y un verbo con sentido completo (como *salir*). En (3) no se señala el movimiento del núcelo (o cabeza) de la proyección de VP hacia el IP, lo que siempre ocurre en el caso de verbos conjugados sin intervención de verbos auxiliares:

4. $[_{IP}\ [_{DP}\ \mathit{pro}]\ \mathbf{sal}_i\mathbf{ió}\ [_{VP}\ [_{V}\ t_i]]]$

Donde *pro* es un sujeto tácito y *t* marca la huella dejada por la raíz verbal *sal-* desde la posición en que es generada hasta aquella en la que es interpretada. La siguiente ilustración nos presenta esta relación, sin considerar el movimiento:

Estructura del sintagma de la inflexión

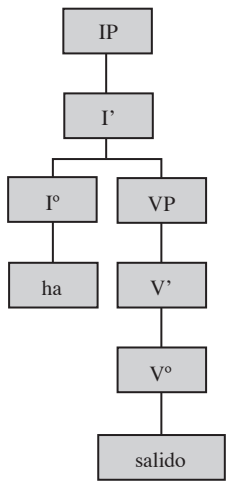

Ilustración 3

Abney (1987) propuso, basado en el hecho de que en varios idiomas las formas nominales contienen morfemas verbales de persona y número, que hay una estructura funcional superior a los NPs, que contiene ragos funcionales (es decir, no referenciales) con los cuales los sustantivos deben concordar. Esta estructura sería paralela a la de la inflexión, como se presenta en (3) y correspondería a la que tenemos en el ejemplo (5) y en la Ilustración 4:

5. $[_{DP}\ [_{D},\ [_{D^\circ}\ \alpha]\ [_{NP}\ \beta]]]$
$[_{DP}\ [_{D},\ [_{D^\circ}\ \mathbf{la}\ ]\ [_{NP}\ \mathbf{casa}\ ]]]$

La nominalización de infinitivos en español puede ilustrar esta relación como lo vemos en (6):

Estructura del sintagma determinante

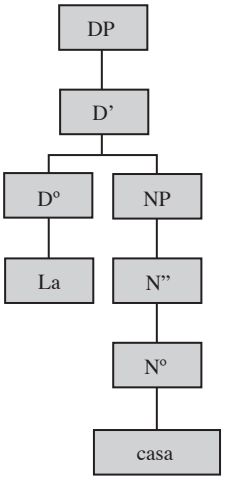

Ilustración 4

Estructrura general del DP

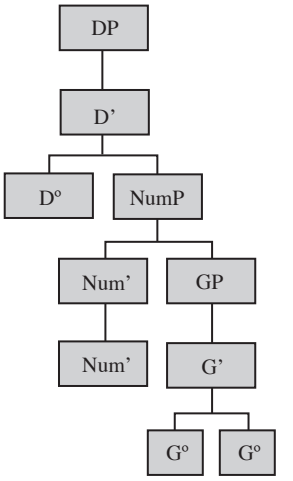

Ilustración 5

6. a. El escuchar …
b. [$_{DP}$ El [$_{CP}$[$_{IP}$ PRO[$_{VP}$ escuchar …]]]]

Donde el DP no recibe ni caso nominativo, ni rol-θ del verbo *escuchar* (es decir, no es ni su sujeto, ni su argumento), sino que contribuye, en esta frase, a nominalizarlo atribuyéndole género (masculino) y número (singular) por cuanto el CP es su argumento; el DP obtiene Caso y rol-θ ya sea siendo sujeto de un IP, o argumento de un VP o un PP, lo cual no se muestra en esta frase. PRO es una categoría vacía que, según la *Teoría de Ligamento*, tiene los rasgos de [+anafórico, +pronominal], en contraste, por ejemplo, con *pro*, como en (4), que es [-anafórico, +pronominal] (es decir, es un pronombre). Por otra parte, la distribución complementaria de los artículos definidos y los pronombres, apoya la idea de que estos pertenecen también a la categoría DP:

7. *El [algún| ningún | aquel | este | mi | tu] libro …

De acuerto con Lorenzo González (1995), es en el nivel de DP donde se dan las operaciones de cuantificación de las formas nominales. La estructura básica del DP en español sería la que tenemos en la ilustración siguiente:

El NumP corresponde al sintagma de número y GP, al de género. De esta estructura se desprende que el núcleo (o cabeza) del NP debe adjuntarse sucesivamente a proyecciones en niveles superiores, moviéndose desde su posición de generación en N°, a G° para obtener género (si no está definido dentro de la información léxica) y a Num° para obtener número; en *Forma Lógica* ascendería a D° para verificar rasgos. Este procedimiento nos concierne en tanto es necesario decidir, en la codificación de NPs, si la entrada contiene una marca de género, en cuya ausencia debe procurársela por este procedimiento sintáctico. Los sustantivos cuyo género está establecido desde el léxico son aquellos de género arbitrario, como *mesa*, *carro*, *radio* o *pared*, los cuales no pueden recibir otro género que el que les corresponde, contrario a términos como *jugador* / *judadora* y *ingeniero* / *ingeniera*.

D'Introno (2002) propone una clasificación de las estructuras nominales, basándose en el criterio de predicación y relativización. De esta forma, establece tres categorías que corresponden a tres niveles de complementación nominal:

1. Argumentales.
2. Argumentales no seleccionados.
3. No argumentales seleccionados.

Los sintagmas nominales se pueden clasificar en dos grupos principales, los argumentales y los no argumentales. Los primeros pertenecen a un nivel mínimo de complementación nominal; los segundos, a un nivel superior de complementación nominal. Esto es demostrable por la posibilidad de movimiento-wh[5] en los sintagmas nominales. Los sintagmas de los grupos 2 y 3 son interpretados como adjuntos al primer elemento predicativo accesible; los complementos argumentales están asociados con los elementos predicativos que los subcategorizan. En la Ilustración 6 se presenta esta clasificación.

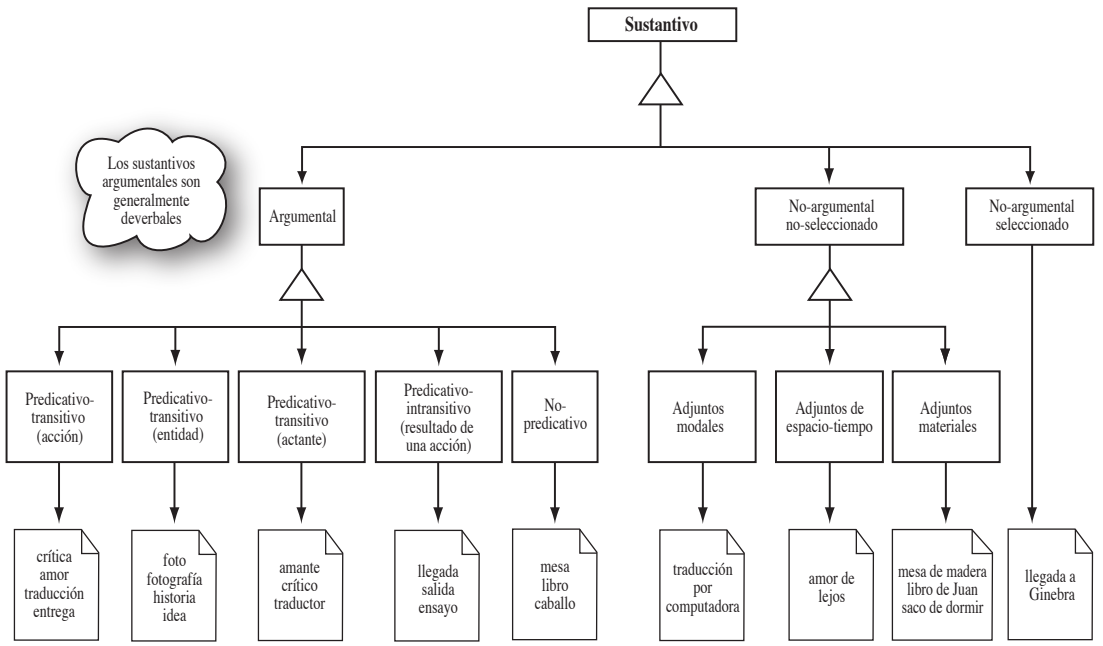

Ilustración 6: Clasificación de las formas nominales según su estructura argumental.

Siendo que el filtro temático (roles-θ) y la estructura de argumentos son el principal criterio de clasificación, nosotros proponemos una reorganización basados en la predicación como criterio delimitador de dos grandes clases, y en la transitividad para las subclases, tal y como se aprecia en la Ilustración 7.

Esta reorganización de los sustantivos privilegia la predicatividad como un rasgo relevante en su estructura léxica; de hecho, todos los NPs predicativos tienen relación con un verbo, sin ser necesariamente deverbales; profundizar en este aspecto implica adentrarse en el universo de las relaciones transcategoriales y del concepto mismo de categoría[6], lo cual esperamos emprender algún día. Todos los sintagmas nominales tienen la capacidad de recibir adjuntos (complementos circunstanciales); sin embargo, en las formas no predicativas, podemos notar que la preposición *de* asigna genitivo:

8. La revista de *la escuela*…
La casa de *mi tío*…

La misma preposición, interactuando con los sustantivos predicativos, se interpreta a partir ya sea de la estructura de argumentos y de los roles-θ (o temáticos), cuando se encuentra al nivel X', o del significado mismo de su argumento:

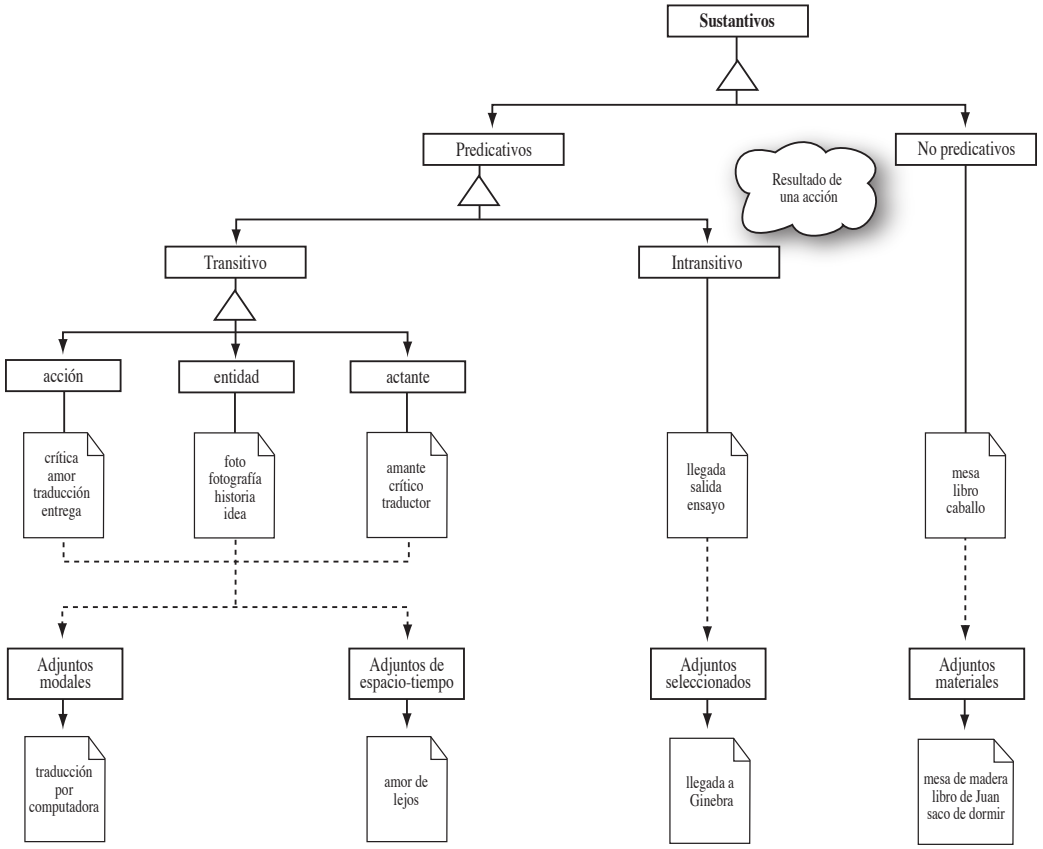

Ilustración 7: Replanteamiento de la clasificación de las formas nominales según los roles temáticos y la estructura de argumentos.

9. Traducción de *la Biblia*…
$[_{DP}\ [_{NP}$ traducción $[_{PP}$ de $[_{DP}$ la Biblia]]]]

10. El cierre de *la escuela* …
$[_{DP}$ El $[_{NP}$ cierre $[_{PP}$ de $[_{DP}$ la escuela]]]]

En (9) y (10) tenemos que *traducción* y *cierre* asignan Acusativo a través de la preposición *de*. La importancia de los roles-θ se puede apreciar claramente si al sustantivo le atribuimos un sujeto:

11. Su traducción de *la Biblia*…
$[_{DP}$ Su$_i$ $[_{NP}$ t_i traducción $[_{PP}$ de $[_{DP}$ la Biblia]]]]

Donde *su* recibe la interpretación de sujeto. Claro está, que otros sentidos no quedan descartados (genitivo por ejemplo), sin embargo, por el filtro temático, la función de sujeto del DP anafórico *su* es preferida. Esto se mantiene en la postnominalización del sujeto:

12. La traducción de Luis de *la Biblia*…
$[_{DP}$ La $[_{NP}$ traducción $[_{PP}$ de $[_{DP}$ Luis]] $[_{PP}$ de $[_{DP}$ la Biblia]]]]

Así como se postula un sujeto interno en VP que recibe el rol-θ correspondiente, para luego desplazarse al SPEC de IP; en (11) tenemos un desplazamiento de un sujeto interno en NP hacia la posición en DP, motivado por Caso (recibe rol-θ, pero no caso) y quizás también por razones fonológicas. En (12) el sujeto permanece en la posición en que es generado, el DP *Luis* recibe rol-θ del NP y Caso del DP. Como lo hemos

visto hasta ahora, la transitividad está relacionada con el nivel X'; en otras palabras, responde a una configuración del árbol sintáctico (argumentos), al Caso Abstracto y a la Teoría-θ. Es precisamente este aspecto el que permite trabajar la desambiguación de las estructuras nominales.

D'Introno (2002) también plantea una estructura casual en la que a través de una serie de movimientos, los núcleos (o cabezas) adquieren, configuracionalmente, Caso Abstracto; esto es otro paralelismo entre los sustantivos y los verbos, los determinantes y la inflexión, que implica una relación de necesidad entre los NPs y ciertas preposiciones, como ya lo hemos visto. No analizaremos en detalle esta vertiente, puesto que en la siguiente sección nos abocaremos a la representación informática de las estructuras nominales planteadas.

3. El tratamiento automático

Por tratamiento entendemos el desarrollo sistemático de una serie de operaciones lógicas y matemáticas efectuadas automáticamente sobre un conjunto de datos para explotarlos en un programa. Si se busca tratar el lenguaje humano, es conveniente recurrir a representaciones formales del funcionamiento del lenguaje, porque facilitan la creación de reglas de operación necesarias para la representación informática de los procesos lingüísticos. En este sentido, la lingüística formal es importante, incluso indispensale para comprender el funcionamiento del lenguaje para establecer sistemas automatizados.

La representación formal que utilizamos está basada en la *Teoría de Rección y Ligamento*, la cual nos ha permitido encontrar relaciones entre las características internas de los lemas y su realización sintáctica. La interpretación informática de dichas relaciones, puede tomar muchas formas. Nosotros escogimos el XML, más adelante expondremos nuestras razones; sin embargo, queremos señalar que existen muchas otras posiblidades, igualmente válidas, cuya escogencia depende de los intereses y de las necesidades del investigador.

3.1. XML

El XML (siglas en inglés del *Lenguaje de Etiquetado Extensible)*[7]; no es un lenguaje de programación, sino más bien, es un sistema de codificación de textos que permite describir contenidos y representar dependencias de un documento. En nuestro caso, se trata de descripciones de unidades léxicas (o lemas). Se lo denomina extensible, porque no tiene un conjunto (pre)determinado de etiquetas (códigos), sino que el investigador las puede crear, siguiendo ciertos principios, según sus necesidades. Las descripciones hechas en XML pueden ser transformadas, por ejemplo, en otros documentos XML (otras representaciones), en texto o en imágenes; nosotros buscamos una manera de utilizarlo para extraer información.

XML describe y establece dependencias, lo que da como resultado que es posible (e incluso recomendable) deducir reglas gramaticales basándose en la codificación misma, lo que se conoce como *esquemas*. El más popular son las *Descripciones de Tipo de Documento* (o DTD), los cuales tratan de describir la estructura de los documentos ; es decir, delinean patrones de las dependencias descritas en XML, gracias a que todo documento XML es representable en forma arborescente. Los DTDs

1. Declaran un conjunto de elementos permitidos y sus atributos.
2. Definen un modelo de contenido para cada elemento.

En síntesis, un DTD establece el conjunto de reglas del lenguaje definido (en nuestro caso se trata de una parcela del lenguaje humano).

El XML puede almacenar y organizar cualquier tipo de información según las necesidades de quien lo requiera. Es un estándar abierto, no está ligado a ninguna compañía o programa. Es aplicable a cualquier sistema de escritura o conjunto de símbolos (runas, ideogramas, alfabeto fonético internacional, etcétera) gracias a que utiliza unicode[8] como su conjunto estándar de caracteres.

Dentro de las características más remarcables del XML se encuentra el hecho de que ofrece muchas maneras de evaluar la calidad del documento con reglas de sintaxis, control interno de vínculos, comparación con modelos de documento (DTDs) y digitación de datos. Además, su estructura es clara, simple y sin ambigüedades; es fácil de leer y analizar tanto por humanos, como por computadoras. En la siguiente ilustración tenemos un ejemplo de un poema codificado en XML (Leoni, 2003):

13. Ejemplo de codificación en XML

```
<?xml version="1.0" encoding="ISO-8859-1"?>
   <poema>
      <titulo>Soneto</titulo>
      <autor>Lope de Vega</autor>
      <estrofa tipo="cuarteto">
            <verso>¿Qué tengo yo que mi amistad procuras?</verso>
            <verso>¿Qué interés se te sigue, Jesús mío?</verso>
            <verso>Que a mi puerta cubierto de rocío</verso>
            <verso>pasas las noches del invierno oscuras.</verso>
      </estrofa>
      <estrofa tipo="cuarteto">
            <verso>¡Oh, cuánto fueron mis entrañas duras!</verso>
            <verso>Pues, no te abrí, ¡qué extraño desvarío!</verso>
            <verso>Si de mi ingratitud el hielo frío</verso>
            <verso>secó las llagas de tus plantas puras.</verso>
      </estrofa>
      <estrofa tipo="terceto">
            <verso>¡Cuántas veces el ángel me decía:</verso>
            <verso>alma asómate agora a la ventana</verso>
            <verso>verás con cuánto amor llamar porfía!</verso>
      </estrofa>
      <estrofa tipo="terceto">
            <verso>¡Y cuántas, hermosura soberana:</verso>
            <verso>"mañana le abriremos", respondía,</verso>
            <verso>para lo mismo responder mañana!</verso>
      </estrofa>
   </poema>
```

Todo documento XML debe comenzar por el prólogo:
<?xml version="1.0" encoding="iso-8859-1" ?>

Este indica:

1. Tipo de documento: se trata de un documento en formato XML.
2. Versión del lenguaje: en este caso es la versión 1.0.
3. Codificación de caracteres: iso-8859-1, es decir, los caracteres están codificados como Latin-1 (alfabetos latinos occidentales).

Otras informaciones que pueden ser declaradas son:

1. DTD utilizado.
2. Declaraciones sobre partes especiales del documento (entidades, por ejemplo).
3. Instrucciones para procesadores de XML.

El resto de la codificación es libre a condición de respetar un cierto número de reglas muy limitado. Todo código debe estar inscrito dentro de los símbolos < y >, donde <> es inicio de código y </> final de código. Es posible crear atributos de etiquetas, los cuales dan un cierto matiz a la etiqueta. En (13) se aprecia que un documento <poema> fue creado. Éste tiene como etiquetas dependientes <autor>, <titulo>[9], <estrofa> y <verso>. A partir de estos podemos dibujar un árbol, que presentamos de manera incompleta:

Árbol parcial de un poema representado en XML

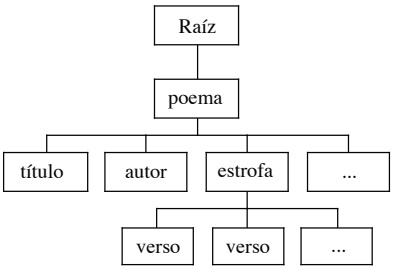

Ilustración 8

Trasladado a estructuras sintácticas, es fácil ver por qué el XML es adaptable a la sintaxis formal: uno de los modos de representación del XML es el árbol, lo cual coincide con nuestra gramática formal, por lo que sus ventajas son múltiples; queremos mencionar que lo más importante para nosotros es la posibilidad de extraer información de manera confiable, además de poder efectuar transformaciones a partir de los datos.

En (14) presentamos un extracto de lo que es nuestra propuesta de codificación, que llamamos *Lenguaje de descripción léxica* (LeXML):

14. Extracto de codificación en LeXML

```
<?xml version="1.0" ?>
<!DOCTYPE Lex SYSTEM "lex.dtd">
<Lex tipo="NP">
<NP predicativo="no">
<Nbarra>
<Nnucleo gen="f">mesa</Nnucleo>
</Nbarra>
</NP>
<NP predicativo="accion">
<Nbarra>
<sujeto rol="agente" />
<PP rol="tema">
de
<DP caso="acusativo" />
</PP>
<Nnucleo gen="f">traducción</Nnucleo>
</Nbarra>
</NP>
<NP predicativo="entidad">
<Nbarra>
<sujeto rol="agente" />
<PP rol="tema">
de
<DP caso="acusativo" />
</PP>
<Nnucleo gen="f">foto</Nnucleo>
</Nbarra>
</NP>
<NP predicativo="entidad">
<Nbarra>
<sujeto rol="paciente" />
<Nnucleo gen="f">foto</Nnucleo>
</Nbarra>
</NP>
<NP predicativo="actante">
<Nbarra>
<PP rol="tema">
de
<DP caso="acusativo" />
</PP>
<Nnucleo gen="f">amante</Nnucleo>
</Nbarra>
</NP>
</Lex>
```

Donde, la categoría sintáctica está implicada en la marca principal <NP> (sustantivo); los atributos de género, rol-θ, predicatividad y caso están señalados respectivamente como *gen*, *rol*, *predicativo* y *caso*; las relaciones de configuracionales y los argumentos se desprenden de la representación arborescente de este documento y los NPs que pueden tener *sujeto* contienen la etiqueta *sujeto*.

En la Ilustración 9, las casillas corresponden a las etiquetas; las que tienen la indicación de *texto* son los nodos terminales que contienen la información. Las líneas punteadas representan los atributos. Este árbol no es exhaustivo, aún para un documento tan pequeño como (14), entre otros elementos, preferimos comenzar por la etiqueta principal <Lex> y no por la raíz del documento. Los puntos suspensivos indican que otros elementos <NP> pueden continuar apareciendo. Un DTD posible para LeXML es el siguiente:

15. DTD básico para LeXML

```
<!ELEMENT NP (#PCDATA | Nbarra)*>
<!ATTLIST NP predicativo CDATA #IMPLIED>
<!ELEMENT PP (#PCDATA | DP)*>
<!ATTLIST PP rol CDATA #IMPLIED>
```

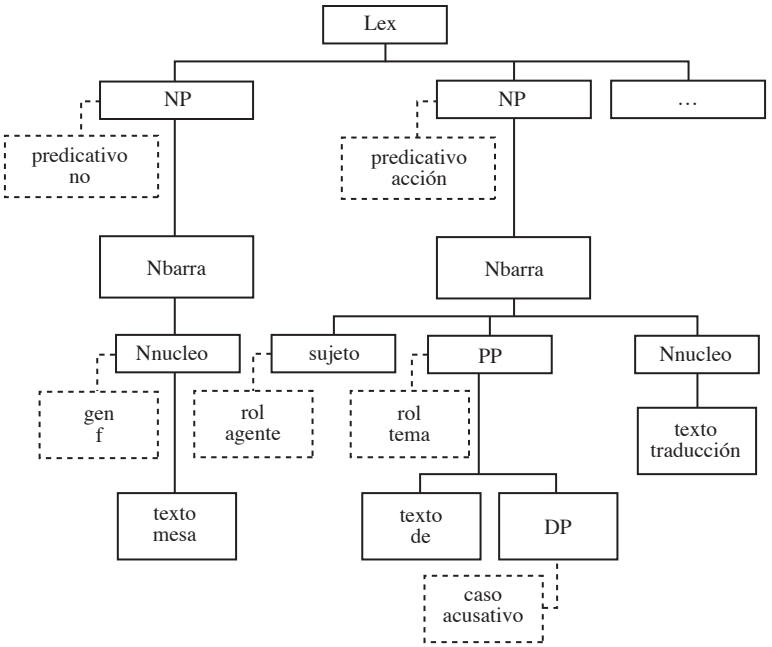

Ilustración 9: Árbol parcial de LeXML

```
<!ELEMENT Lex (#PCDATA | NP)*>
<!ATTLIST Lex tipo CDATA #IMPLIED>
<!ELEMENT Nbarra (#PCDATA | Nnucleo | sujeto | PP)*>
<!ELEMENT Nnucleo (#PCDATA)>
<!ATTLIST Nnucleo gen CDATA #IMPLIED>
<!ELEMENT DP (#PCDATA)>
<!ATTLIST DP caso CDATA #IMPLIED>
<!ELEMENT sujeto (#PCDATA)>
<!ATTLIST sujeto rol CDATA #IMPLIED>
```

Entre paréntesis se indican las marcas dependientes. La barra vertical "|" equivale a la conjunción disyuntiva "*o*", el asterisco "*" significa uno o más de los elementos precedentes, el signo de suma "+" *uno o más* y el signo interrogativo de cierre, *cero o uno*. Esta es sólo una de las opciones para el DTD, cuyo propósito es impedir la creación de documentos XML incorrectos, es decir, que no correspondan a las reglas de la gramática.

3.2. Extracción y transformación de datos

Para extraer y transformar datos, una opción es el *Lenguaje de transformaciones de hojas de estilo extensible* (XSLT por sus siglas en inglés). Es una herramienta de publicación de contenidos estructurados dentro del formato del *Lenguaje de etiquetado extensible* (XML).

El XSLT permite extraer contenidos de un documento XML y verterlos en otros formatos (incluyendo XML). Por ejemplo, es posible escoger los elementos de un documento XML y crear con ellos un documento HTML, con la indicación exacta de las características de presentación y de los entornos en que deben aparecer..[10] En nuestro caso particular, es posible tomar crear, a partir de una base de datos en XML que representa el léxico, formar una serie de documentos XML que figuren frases creadas a partir de los datos léxicos. Por supuesto que esto implica un trabajo mucho más grande que el que hemos hecho hasta aquí, pero es un ejemplo de un uso concreto.

Un programa XSLT es un documento XML con instrucciones de lectura y procesamiento de otro documento XML, que también incluye un modelo del documento que se creará y en el que deberá ser vertido el contenido. El programa aprovecha la representación arborescente

realizada por un procesador[11] de XSLT para orientarse en el documento XML como si se tratara de un mapa. Cada sección del documento original en XML es identificada por el nombre de la etiqueta (el elemento) y sus atributos, siguiendo esquemas configuracionales descritos en el *Lenguaje de caminos* (XPath). Una vez identificada y extraída la información, esta es vertida en el modelo y así se crea el documento meta.

Como ejemplo, en la *hoja XSLT* en (16) tenemos un programa XSLT que extrae únicamente los lemas marcados como "*predicativo='no'*" y los inserta en un documento HTML sencillo:

16. Extracción de elementos no predicativos en LeXML con XSLT

```
<?xml version="1.0" encoding="UTF-8" ?>
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.
    w3.org/1999/XSL/Transform">
<xsl:template match="/">
<html>
<head>
<title>
<xsl:text>Rasgo predicativo="no"</xsl:text>
</title>
<link rel="stylesheet" type="text/css" href="AdjML.css"
    />
</head>
<body>
<h1>
<xsl:text>Predicativo = "no"</xsl:text>
</h1>
<ol>
<xsl:apply-templates select="/Lex/NP[@predicativo='no']"
    />
</ol>
</body>
</html>
</xsl:template>

<xsl:template match="Nnucleo">
<li>
<xsl:value-of select="text()" />
</li>
</xsl:template>
</xsl:stylesheet>
```

El resultado es puesto en un documento HTML que sólo consiste del dato "mesa" por cuanto es el único lema no predicativo de nuestra pequeña base de datos. En contextos de uso masivo de esta representación léxica, se puede recurrir a otros lenguajes informáticos como Java o Perl, con los cuales es posible crear ambientes dinámicos completaos en la que los usuarios interactúan continuamente con la base de datos (directa o indirectamente). En (1) tenemos un ejemplo de extracción de datos utilizando el procesador *Xalan*:

17. Extracción de datos con Xalan en una sesión en Unix:

```
[comp7:xalan/samples/ApplyXPath] usuario10%
    java ApplyXPath npxml-2.xml "/Lex/NP[@
    predicativo='actante']/Nbarra/Nnucleo"
Loading classes, parsing npxml-2.xml, and setting up seria-
    lizer
Querying DOM using /Lex/NP[@predicativo='actante']/
    Nbarra/Nnucleo
<output>
<Nnucleo gen="f">amante</Nnucleo>
</output>
[comp7:xalan/samples/ApplyXPath] usuario10%
```

En (17) se busca un lema con el rasgo "*predicativo=actante*"en la base de datos, a partir de la línea de comando (o terminal), utilizando una expresión del *Lenguaje de caminos* (XPath), el cual describe relaciones configuracionales en un árbol XML. En (18) tenemos otro uso de *Xalan* con expresiones *XPath*:

18. Extracción de un N° que asigna acusativo

```
[comp7:xalan/samples/ApplyXPath] usuario10%
    java ApplyXPath npxml-2.xml "/Lex/NP[@
    predicativo='actante']/Nbarra/Nnucleo/../PP/DP[@
    caso='acusativo']/../../Nnucleo"
Loading classes, parsing npxml-2.xml, and setting up
    serializer
Querying DOM using /Lex/NP[@predicativo='actante']/
    Nbarra/Nnucleo/../PP/DP[@caso='acusativo']/../../
    Nnucleo
<output>
<Nnucleo gen="f">amante</Nnucleo>
</output>
[comp7:xalan/samples/ApplyXPath] usuario10%
```

En (18) se muestra cómo se puede aprovechar la información configuracional para extraer datos. En este caso, la idea era obtener un N° con el rasgo *predicativo='actante'* que asignara acusativo, lo cual hace a través de un PP. En nuestra base de datos, el único lema que cumple estos requisitos es *amante*. La inversión por medio de indicación de ascenso de nivel (`/../../Nnucleo`) se debe a que queremos recuperar núclo (o cabeza) y no el argumento como habría sido el caso sin tal instrucción.

Es esta facilidad para describir y recuperar datos, lo que hace posible utilizar cualquier documento XML para intercambiar datos, así como para utilizarlos en diversas aplicaciones. LeXML no es la excepción, pudiéndose recurrir a él para intercambiar datos léxicos (memoras de traducción, actualización de diccionarios, etcétera). Por otra parte, también es posible integrar este tipo de documentos en bases de datos. En los últimos años ha habido un esfuerzo muy grande por crear sistemas que hagan factible el uso del XML con grandes cantidades de información con tiempos de consulta y actualización aceptables sin que las ventajas intrínsecas al XML se pierdan (representación compleja de datos en forma de estructuras interdependientes). Este es un tema por desarrollar y que sin duda en el área de las Letras podrá ser de gran beneficio, dada la naturaleza de los datos en humanidades.

4. Conclusiones

La noción del DP como estructura paralela al IP, y de los sintagmas nominales como correspondientes a los verbales, enriquece las posibilidades de análisis del léxico, desde una perspectiva sintáctica, por cuanto se evidencia la trama de relaciones que median entre los NPs y su realización en el enunciado.

Precisamente, una de los resultados de tal acercamiento, es posibilidad de plantear una desambiguación, no determinista, desde las formas nominales almacenadas en el léxico, a partir de su estructura de argumentos, roles temáticos y Caso.

El XML es una herramienta con grandes potencialidades en la Lingüística, por cuanto permite codificar estructuras complejas que pueden ser explotadas por medios informáticos de extracción de datos. Su incorporación dentro de las instrumentos para análisis automáticos del lenguaje es deseable por cuanto permite reproducir las estructuras lingüísticas, casi de la misma forma en que son admitidas en la investigación.

El análisis del léxico está lejos de ser concluido, más bien consideramos que hemos trazado un sendero que merece ser recorrido y que, a la vez, es una invitación para que otros se decidan a recorrerlo.

Notas

1 Wehrli (1997:25).

2 Ver Schönfeld (2001: 6-13, 123- 132)

3 Estos temas, en relación con las estructuras sintácticas, pueden ser objeto de otro tema de investigación.

4 Utilizamos la nomenclatura inglesa para designar los sintagmas, de esta forma DP es el sintagma determinante; NP, el nominal; VP, el verbal; IP, la inflexión, etcétera.

5 Es decir, de relativización.

6 Cabe mencionar también, que incluso la función msima de los morfemas y la existencia de lemas como entidades abstractas serían parte de tal investigación.

7 En inglés: *Extensible Markup Language*.

8 Unicode es una propuesta de representación universal de caracteres, de tal manera que sea posible utilizar, por ejemplo, tanto el chino, como el español, dentro de un mismo formato. Su objetivo es acabar con el exceso de formatos de caracteres que actualmente existe.

9 Por cuestiones metodológicas y de compatibilidad de sistemas, se recomienda utlizar únicamente los caracteres ASCII en las etiquetas; en nuestro caso eso significa prescindir de los caracteres acentuados.

10 Por ejemplo en tablas, entre etiquetas de negritas o de hipervínculos.

11 Algunos de los procesadores más conocidos son Xalan, Xerces, libxml2 y libxslt, xmllint y xsltproc. Estos son gratuitos, aunque ambién hay opciones comerciales con licencias de uso y contratos de atención al cliente. Algunas direcciones útiles son: http://xml.apache.org y http.//www.xmlsoft.org.

Bibliografía

Abney, Steven. 1987. The English NP in its sentential aspect. Tesis doctoral, MIT.

Amann, Bernd y Philippe Rigaux. 2000. *Comprendre XSLT*. O'Reilly & Associates, París.

Borer, Hagit. 1984. *Parametric Syntax*. Dordrecht: Foris Publications.

Bray, Tim et al. *El lenguaje extensible de marcas* (XML) 1.0 - REC-xml-19980210 - [página de internet] febrero de 1998; http://www.thefaactory.com/ta/xmlspec/index.html [5 de enero de 2002].

Bray, Tim et al. *Espacios de nombre en XML* - [página de internet] 14 de enero de 1999; http://html.conclase.net/w3c/xml-names-es/ [12 de junio de 2003].

Bray, Tim et al. *Espacios de nombre en XML* - [página de internet] 14 de enero de 1999; http://html.conclase.net/w3c/xml-names-es/ [12 de junio de 2003].

Bray, Tim et al. *Extensible Markup Language (XML) 1.0 - El lenguaje extensible de marcas (XML) 1.0* - [página de internet] febrero de 1998; http://www.sidar.org/recur/desdi/traduc/es/xml/xml1/index.html [12 de junio de 2003].

Brody, Michael. *Lexico-logical form*. Cambridge, Mass.: MIT Press.

Burzio, Luigi. 1986. *Italian Syntax: A Government-Binding Approach*. D.Reidel Publishing Company, Dordrecht.

Chametzky, Robert A. 2000. *Phrase Structure from GB to Minimalism*. Malden, Mass.: Blackwell Publishers.

Chomsky, Noam. 1973. *Conditions on Transformations*. En: S. Anderson and P. Kiparsky (eds.), *A Festschrift for morris Halle*. New York: Holt, Reinhart and Winston.

Chomsky, Noam. 1981. *Lectures on Binding and Government*. Dordrecht: Foris Publications.

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

Clark, James y Steve de Rose. *Lenguaje de caminos XML (XPath)* - [página de internet] 19 de octubre de 2001; http://www.sidar.org/recur/desdi/traduc/es/xml/xpath.html [12 de junio de 2001].

Cover, Robin. XML Cover Pages: Academic Applications - [página de internet] 10 de mayo 2001; http://www.oasis-open.org/cover/acadapps.html [12 de diciembre de 2001].

Culicover, P. W. 1997. *Principles and Parameters: An Introduction to Syntactic Theory*. Oxford University Press, Oxford.

D'Introno, Francesco . 2001. *Sintaxis Generativa del Español: Evolución y Análisis*. Madrid: Cátedra.

D'Introno, Francesco. 2002. *Niveles de complementación nominal*. Sin publicar. Caracas:

Enç, Mürvet. 1991. «The Semantics of Specificity», *Linguistic Inquiry* 22: p. 1-25; MIT Press.

González, Guillermo Lorenzo (1995. Geometría de las estructuras nominales: sintaxis y semántica del SDET. Oviedo: Departamento de Filología Española.

Grimshaw, Jane. 1990. *Argument Structure*. Cambridge, Mass.: MIT Press.

Haegeman, Liliane. 1994. *Introduction to Government & Binding Theory*. Blackwell Publishers, Oxford.

Heim, Irene y Angelika Kratzer (1998. *Semantics in Generative Grammar*. Malden, Mass.: Blackwell Publishers

Hernanz, Ma. Lluïsa y José Ma. Brucart. 1987. *La Sintaxis: 1. Principios teóricos. La oración simple*. Barcelona: Editorial Crítica.

Kayne, R. 1984. *Connectedness and Binary Branching*. Foris, Dordrecht.

Lehrer, Adrienne. 1974. *Semantic fields and Lexical Structure*. Amsterdam: American Elsevier.

Leoni de León, Jorge Antonio. 2003. *Propuesta de formalización del manual de redacción del Nuevo Diccionario del Español de Costa Rica a partir del Lenguaje de Etiquetado Extensible (XML)*. Káñina, Revista de Artes y Letras, Universidad de Costa Rica. Vol.XXVII (2), 2003.

May, Robert. 1985. *Logical Form*. Cambridge, Mass.: MIT Press.

Miriam, Butt y Wilhelm Gender. 1998. *The projection of arguments: lexical and compositional factor*. Stanford: CSLI Publications.

Patrick Saint Dizier y Evelyne Viegas, eds. 1995. *Computational lexical semantics*. Cambridge, UK: Cambridge University Press.

Pesetsky, David. 1995. *Zero Syntax*. Cambridge, Mass.: MIT Press.

Radford, Andrew. 1997. *Syntax: A minimalist introduction*. Cambridge University Press, Cambridge.

Ravin, Yael. 1990. *Lexical semantics without thematic roles*. Oxford: Oxford University Press.

Ray, Erik. 2001. *Learning XML*.O'Reilly: California.

RIZZI, Luigi. 1982. *Issues in Italian Syntax*. Foris, Dordrecht.

RIZZI, Luigi. 1986. *On chain formation*. Dans Borer, Hagit (ed.), *Syntax and Semantics*, pp. 65-95, Academic Press, Inc., Orlando.

Rizzi, Luigi. 1990. *Relativized Minimality*. Cambridge, Mass.: MIT Press.

Rosen, Sara Thomas. 1990. *Argument Structure and complex predicates*. New York: Garland Publishers.

Schönefeld, Doris. 2001. *Where Lexicon and Syntax meet*. Berlin – New York : Mouton de Gruyter.

SIGLEX Workshop. 1st 1991. *Lexical semantics and knowledge representation*. First SIGLEX.

Speas, Margaret J. 1990. *Phrase Structure in Natural Language*. Dordrecht/Boston/ London: Kluwer Academic Publishers.

Steven L. Small et al. 1988. *Lexical ambiguity resolution perspectives from psycholinguistics*. Morgan Kaufmann Publishers.

The World Wide Web Consortium. 2001. *The World Wide Web Consortium* - [página de internet] 12 de diciembre de 2001; http:// www.w3.org/ [12 de diciembre de 2001].

Tim Bray et al. 1998. *El lenguaje extensible de marcas* (XML) 1.0 - REC-xml-19980210 - [página de internet] febrero de 1998; http:// www.thefaactory.com/ta/xmlspec/index. html [5 de enero de 2002].

WEHRLI, Eric. 1997. *L'Analyse Syntaxique des Langues Naturelles: Problèmes et Méthodes*. Masson, Paris.

William, Edwin. *Thematic structure in Syntax*. Cambridge, Mass.: MIT Press.

Yorick A. Wilds, Brian M. Slator y Louise M. Guthrie. 1996. *Electric Words: Dictionaries, Computers, and Meanings*. Cambridge, Mass. : MIT Press.

Zubizarreta, María Luisa. 1987. *Levels of representation in the lexicon and in the syntax*. Dordrecht/ Providence, RI: Foris Publications.