PRÓGENES: LA INTERFAZ EN LENGUAJE NATURAL

Julia Díaz García.
Pilar Rodríguez Marín.

Descripción del problema PRÓGENES

El objetivo de este artículo es describir el módulo de tratamiento del lenguaje natural que se va a utilizar dentro del proyecto de deducción automática PRÓGENES (PROof GENerator Expert System).

PRÓGENES comenzó a principios de 1990 y su duración estimada es de unos tres años. Al cabo de este tiempo se dispondrá de un sistema de deducción completamente automática, capaz de resolver por sí mismo los problemas de matemáticas que se le planteen, correspondientes a la materia impartida en el primer curso de las carreras de ciencias. En este sentido, PRÓGENES se examinará de la asignatura junto con los alumnos del curso correspondiente.

En un principio se podría suponer que la entrada del sistema fuesen los enunciados de los problemas expresados en un lenguaje formal determinado, pero el objetivo es más ambicioso, ya que se trata de proporcionarle exclusivamente la misma información, y en la misma forma, que a los alumnos. Por lo tanto, se ha incorporado al sistema una interfaz capaz de entender el texto de los enunciados expresados en lenguaje natural, español, e interpretarlos según una semántica que refleja los conceptos y relaciones existentes en esta área de conocimientos.

Con este planteamiento, la interfaz se convierte en un elemento clave para el buen fin del sistema, ya que la resolución de los problemas tiene lugar a partir de la expresión semántica que se le suministra y que, a su vez, es el resultado del análisis del texto origen.

Características

El dominio en el que nos movemos es en principio restringido, centrándose en el lenguaje utilizado para el planteamiento de los problemas de matemáticas de un primer curso de carrera universitaria de ciencias. Como cualquier lenguaje especializado, este tipo de textos tiene unas características peculiares cuyo análisis determina, principalmente, algunos problemas a los que hay que prestar gran atención, dado que partimos de la base de que el resultado final de este proceso es una expresión formal del contenido del enunciado absolutamente libre de ambigüedades.

Dada la restricción del dominio, es de prever que no nos encontremos con grandes problemas de ambigüedad léxica. Sin embargo, en lo relativo al discurso la situación es muy diferente.Este problema,entendido como la resolución de referentes, es siempre muy importante, pero en algunas aplicaciones se evita a menudo, asumiendo que la indeterminación que se introduce no es crítica de cara al resultado final. Nos encontramos con esta situación, por ejemplo, en algunas aplicaciones de traducción automática.

En PROGENES no podemos asumir esta indeterminación ya que para la correcta formalización del enunciado han de considerarse todos los datos que éste especifique. El enunciado puede constar de una o varias líneas, y puede incluir subapartados, notas u otros tipos de descripciones. De manera que es necesario resolver, por una parte,las referencias dentro del alcance de cada oración aislada (pronombres, subordinadas, etc.) y, por la otra, las referencias que se den en un contexto formado por varias oraciones.

Este último punto incluye tanto las referencias que se realizan mediante la utilización explícita de pronombres, como la interpretación conjunta de las facciones de información que se encuentran en cada una de las oraciones del enunciado:

(1) "Hallar el punto donde se intersectan las rectas 2x-7y+31=0 y x-2y+7=0, y determinar su ángulo de intersección."

(2) "Sean a y b números no negativos. Demostrar que si $a^2 =< b^2$ entonces a =< b.
INDICACION: $b^2-a^2=(b+a)(b-a)$."

(3) "Escribir una ecuación para la recta l_2 paralela a 3x-5y+8=0 y que pasa por el punto (-3,2)."

Por ejemplo, en (1) es absolutamente necesario determinar que "*su ángulo de intersección*" se refiere a *"las rectas"*, evitando relacionarlo con *"el punto"*. A su vez, el enunciado (2) consta de tres oraciones que proporcionan los datos y plantean el problema. En ambos casos la interpretación del texto debe tener en cuenta todas las relaciones existentes: la referencia a la recta en (1) y los términos *"a"* y *"b"* que hacen referencia a las mismas entidades en las distintas porciones de (2). Igual que en (1) y (2), en (3) se pueden encontrar casos de referencias que no son inmediatas.

En este sentido, también es crítico para el sistema el problema de la determinación del nivel de modificación al que se refieren los complementos preposicionales, aún siendo relativamente tolerable en muchas otras aplicaciones. Son frecuentes situaciones como la del ejemplo siguiente, donde es absolutamente necesario determinar la correcta asociación de los complementos preposicionales:

(4) "Hallar la intersección de la recta R con el eje X."

Para la asociación del primero de los dos grupos preposicionales que aparecen no existe ningún problema, ya que sólo es posible que dependa de *"intersección"* (la dependencia de *"Hallar"* es fácilmente descartable). Igualmente, *"con el eje X"* tiene que asociarse a *"intersección"* y no a *"recta"*:

(4-1) "la intersección de A con B"
*(4-2) "la recta con C"

(4-1) tiene sentido, siempre y cuando se cumplan algunas restricciones sobre el ***tipo*** de A y de B que, como veremos a continuación, también hay que considerar en el sistema.

Concretamente, A y B han de ser de tipo ***curva***, circunstancia que sí se cumple en la oración (4). (4-2) pudiera tener algún sentido, pero no de forma evidente en nuestro dominio.

Arquitectura

A partir de los requerimientos, se ha diseñado una arquitectura que consta de dos partes en principio bien diferenciadas: 1) la interfaz para la interpretación del enunciado y 2) el módulo de deducción. Estas dos partes son básicamente disjuntas en cuanto a su funcionamiento, pero han de compartir el conocimiento del dominio al que se aplica el sistema. Este conocimiento se representa en dos niveles diferentes, pero relacionados [MCS91]:

- Por una parte se ha determinado el grupo de *tipos* válidos en el entorno, como son:

(5)
RECTA
EJE
PUNTO
ACCION
...

y que, en conjunto, constituyen una jerarquía unidireccional.

(6) EJE : RECTA : CURVA : ...

Este conocimiento está siempre disponible, tanto para el módulo de tratamiento del lenguaje, como para el de deducción.

- Por otra parte, se describen todos los conceptos que van a contemplarse, incluyéndose, para cada uno de ellos, una información que se considera indispensable:
- *Tipo*: uno de los tipos válidos de los que hemos hablado en el punto anterior (ver ejemplo (5)).
- *Descriptores*: dado un concepto, se indican bajo este nombre todos los elementos relacionados con él, que a su vez son conceptos cuyo tipo ha de aparecer explícitamente.

Son expresiones válidas:

(7) (**hallar** lista(OBJETO) / usando : lista(BOOL)) : ACCION
(**recta** ecuación : lista(ECUACION) /
dimambiente : NUMPROG sistcoord : SISTCOORD
FORMA-ECUACION) : RECTA

donde la primera palabra de cada entrada indica el concepto que se está describiendo (**hallar**, **recta**,...), y cuyo tipo es la clave que se encuentra al final de la expresión y separada por "**:**". Entre el valor del concepto y el tipo asignado aparecen los descriptores, que determinan cuales son aquellos datos o conceptos relacionados con aquel que se está describiendo. Estos descriptores pueden estar precedidos por el símbolo "/", significando entonces que su ocurrencia es opcional.

También se da el caso de que algunos descriptores tengan asociado un nombre, como *ecuación* en el caso de recta, y otros no. Se da un nombre al descriptor cuando éste es a su vez un concepto cuya descripción está implícita. En la definición de **recta** está implícito el concepto **ecuación**, por lo que es deducible y válida la siguiente expresión semántica:

**** figura 1 ****

(8) (**ecuación** RECTA) : lista(ECUACION)

Desde el punto de vista del procesamiento del lenguaje natural, la base de conocimientos que contiene los conceptos matemáticos no es directamente asimilable a un diccionario semántico. Antes de poder ser utilizada es necesario establecer una equivalencia entre las palabras del léxico y los conceptos que significan. Esta equivalencia es de muchos a uno en varios casos, ya que, a pesar de lo restringido del entorno, se ha detectado que se utilizan varias expresiones diferentes para significar la misma idea.

(9) "Escribir la ecuación..." -> (**hallar** ...)
"Hallar la ecuación..." -> (**hallar** ...)

También se da el caso de equivalencias de uno a muchos, como sucede con el ejemplo (8), **ecuación**, que se puede deducir a partir de la definición de recta en (7), pero que también se puede obtener a partir de la de plano y de la de otros muchos objetos matemáticos, determinándose el sentido correcto durante el análisis.

(10) (**recta** ecuación : lista(ECUACION) ...)
=> (**ecuación** RECTA) : lista(ECUACION)

(11) (**plano** ecuación : ECUACION ...)
=> (**ecuación** PLANO) : ECUACION

Como se presenta en la figura anterior, tenemos que el diccionario PRÓGENES se construye a partir de la base de conocimientos semántica (conceptos de la materia) y del léxico propiamente dicho que describe el comportamiento de las palabras. La aproximación lingüística es básicamente lexicalista, como detallamos a continuación, por lo que el diccionario resulta ser el principal componente del sistema.

El texto introducido es el del enunciado completo de cada problema a resolver que es analizado hasta llegar a una estructura que representa el enunciado. Esta estructura incluye una expresión semántica que es convertida de forma casi inmediata a una expresión PRÓGENES válida, siendo expresiones de estas características todas las que verifican la sintaxis determinada por la base de conocimientos semántica del principio, y que da lugar al lenguaje formal PRÓGENES. Este lenguaje, junto con la jerarquía de tipos y algunas variables de entorno, constituyen la única forma de comunicación compartida entre la interfaz y el módulo de deducción automática.

Estrategia lingüística

En la fase de diseño de la interfaz se decidió abordar el problema del procesamiento del lenguaje con el que nos enfrentábamos mediante un formalismo de gramática categorial de unificación (Unification Categorial Grammar), después de barajar otras posibilidades (ATNs, gramáticas basadas en reglas...).

Esta decisión está apoyada, en primer lugar, porque pensamos que es conveniente hacer recaer la mayor parte del peso del proceso sobre el léxico, tendencia cada vez más admitida. En ese sentido las aproximaciones basadas en UCG se apoyan fundamentalmente en la descripción de cada ocurrencia de las palabras, descripciones que incluso incluyen la categoría sintáctica de la palabra expresada de forma dinámica (ver el apartado siguiente) [KAY86] [REY88]. Por otra parte, la implementación de UCG es adecuada desde el punto de vista computacional, obteniéndose buenos resultados en cuanto a la rapidez del proceso [BIH88], [WIW90], [YIK90]. Por último nos decidió la gran importancia que tiene la extracción semántica en nuestra aplicación, extracción que es inmediata en las aproximaciones basadas en UCG [CKZ88].

El diccionario

Sistematizando la creación del diccionario se han definido un conjunto de rasgos básicos formados por pares atributo-valor, o atributo-estructura, para describir cada entrada:

base Cuyo valor puede ser:

- La forma canónica de la palabra, i.e., el infinitivo de los verbos y la forma masculina y singular de los nombres, adjetivos y determinantes. Las preposiciones, adverbios y conjunciones se mantienen invariantes.
- La composición de formas canónicas que consiste en la secuencia de éstas separadas por el signo "+".

cat Su categoría sintáctica asociada.

núcleo donde se incluyen bajo el nombre *concordancia* los rasgos morfológicos: género, número y persona y bajo el nombre *sem* la descripción semántica del núcleo en aquellos casos en los que sea necesario.

sem Su expresión semántica.

Los atributos **núcleo** y **sem** tienen asignados valores por defecto que se incluyen en aquellas palabras que, en nuestro dominio, no tienen semántica explícita, como ocurre con las preposiciones, adverbios y conjunciones.Unicamente durante el análisis y mediante unificación se instancian con valores específicos.

El diccionario se ha implementado con CLOS (Common Lisp Object System) considerando una única entrada con los valores por defecto descritos en el párrafo anterior y creando cada palabra como instancia de la entrada general, estableciendo siempre la herencia de los atributos y, cuando sea necesario, la herencia de los valores o estructuras asociadas.

La representación semántica del lenguaje se realiza partiendo de la definición de la palabra en el léxico PRÓGENES, que se encuentra en la base de conocimientos de conceptos, y teniendo en cuenta los siguientes criterios [CKZ88]:

- Cualquier fórmula semántica va precedida por un índice, que formamos con la(s) inicial(es) de su tipo correspondiente. Este tipo corresponde a un tipo válido en la jerarquía.

- A continuación del tipo aparece el ***concepto*** seguido de una lista de argumentos cuyos elementos son las expresiones semánticas de sus ***descriptores*** asociados. La semántica de los descriptores debe poder deducirse de los elementos sobre los que se aplican los operadores "/" y "\". En otro caso, la expresión semántica global aparece dentro de una secuencia entre corchetes seguida del comodín *S*.

Para ilustrar la creación de las expresiones semánticas recordemos el concepto **recta** que vimos anteriormente. Obviando los descriptores opcionales, tenemos:

(7') [r] **RECTA** ([e] @E)

denotando [r] y [e] los tipos RECTA y ECUACION respectivamente, y representando @E a la ecuación necesaria para identificar la recta.

Presentamos otro ejemplo más con la palabra "*paralelo*", que en el lenguaje formal PRÓGENES se describe como en (11), y cuya expresión semántica resulta (11'):

(11) (**paralelo** [VECTOR|RECTA|PLANO] [VECTOR|RECTA|PLANO]) : BOOL

(11') [b] PARALELO ([v|r|p] @P , [v|r|p] @Q)

siendo [b] BOOLEANO, [v] VECTOR, [r] RECTA y [p] PLANO.

En ambas definiciones, las líneas verticales "|" indican una disyunción de valores, ya que un vector (idem para recta o plano) puede ser paralelo a otro vector, a una recta o a un plano. Análogamente al caso anterior, @P y @Q son descriptores necesarios para la determinación del concepto **paralelo**: *"(la recta) @P es paralela a (la recta) @Q"*.

En definitiva, tenemos las entradas léxicas para las palabras *recta* y *paralela* que se representan a continuación, tal como necesitamos recuperarlas en la expresión *"... recta paralela a ..."*. En ellas se reune la información sintáctica sobre su categoría junto con la descripción semántica y el resto de los rasgos que mencionamos anteriormente.

*** figura 2 ****

La Gramática

Como hemos visto, nuestra aproximación se caracteriza por incluir la mayor parte de la información en el diccionario, como es habitual en UCG, utilizándose muy pocas reglas combinatorias que, además, sólo afectan a las categorías asignadas en el léxico, y permitiéndose una única operación sobre los objetos gramaticales, que es la unificación [SHI86]. En parte por lo anterior, las expresiones sintácticas y semánticas asociadas a las palabras están muy relacionadas.

En este marco, UCG combina los puntos de vista sintácticos de las Gramáticas Categoriales con los puntos de vista semánticos de la Teoría de la Representación del Discurso, siendo la potencia del conjunto aún mayor al utilizarse mecanismos de unificación que permiten relacionar diferentes niveles lingüísticos en la descripción de las palabras.

Clásicamente, se emplean tres categorías primitivas: **N**, **S** y **NP**. En nuestra implementación hasta ahora sólo hemos utilizado las dos últimas, no siéndonos necesario por el momento considerar *N* como categoría primitiva.

El resto de las categorías se obtiene como sigue:

1. Cualquier categoría primitiva es una categoría.
2. Si A y B son categorías, también lo es A/B y A\B.

Sobre estas categorías se definen las operaciones clásicas que constituyen la gramática combinatoria, usando la notación de Steedman [WIW90], [MIS91]:

_ X/Y	Y	_	X	**fa**
_ Y	X\Y	_	X	**ba**
_ X/Y	Y/Z	_	X/Z	**fc**
_ Y\Z	X\Y	_	X\Z	**bc**
_ X _ Y / (Y \ X)				**tr**
_ X _ Y \ (Y / X)				**tr**

Aunque el análisis morfológico, sintáctico y semántico se realizan simultáneamente, a continuación describimos más en detalle el análisis sintáctico de un fragmento del enunciado (3), para ilustrar la aplicación de la gramática sobre la descripción de las palabras:

(12) que pasa por el punto $P(-3,2)$

Cada palabra tiene asociada una categoría sintáctica, que aparece debajo. Por la acción de las reglas indicadas anteriormente, las estructuras unitarias se combinan de la siguiente manera:

**** figura 5 ****

Las posibilidades combinatorias de la aplicación funcional pueden producir más árboles con análisis válidos, pero las restricciones impuestas por la semántica deciden su unicidad. Se puede observar su precisión y economía motivada por el carácter cuasi-aritmético de las reglas.

Un ejemplo

Los exámenes de matemáticas que el sistema resolverá son los de primer curso de ciencias de la Universidad Autónoma de Madrid. Estos exámenes se escriben habitualmente utilizando un procesador de texto en concreto, que resulta ser TeX. A los alumnos se les suministra una copia del examen en papel, una vez compilado. A PRÓGENES se le va a suministrar exactamente el mismo enunciado, pero tal y como está escrito antes de ser procesado. Por ejemplo:

(13) \#1.3.1. Escribir una ecuación para la recta paralela a

$$

3x-5y+8=0
$$

y que pasa por el punto $P(-3,2)$.

En un primer momento, el texto de entrada es estudiado para determinar cuales de sus elementos son palabras y cuales son marcas, fórmulas o expresiones. Las expresiones, una vez identificadas, son tratadas por un módulo analizador que determina sus características, y es capaz de clasificarlas en uno de los tipos posibles de la jerarquía. Este tratamiento ayuda a determinar su estructura léxica para el análisis posterior. En el ejemplo anterior, la expresión *3x-5y+8=0* una vez analizada dará lugar a las siguientes entradas léxicas:

**** figura 3 ***

Suponiendo, que el valor de la dimensión ambiente es 2. En dimensión 3 la segunda entrada sería un plano. Igual sucede con la estructura de *(-3,2)*. El resto de las palabras son buscadas en el diccionario PRÓGENES, donde existen una o varias entradas para cada una de ellas como se ha descrito en los apartados anteriores. Estas entradas todavía no son definitivas y responden a las características de las palabras con las que hemos trabajado hasta el momento. Unificando de forma binaria las entradas obtenidas del diccionario llegamos al resultado final del análisis:

**** figura 4 ***

donde a partir del valor del atributo **sem** se obtiene de manera casi inmediata la traducción del enunciado (13) al lenguaje formal PRÓGENES.

Conclusión

Desde el punto de vista lingüístico, uno de nuestros objetivos era comprobar la adecuación de una teoría tan lexicalista como UCG para la aproximación al procesamiento de un lenguaje fuertemente dependiente del dominio como es el que se maneja en este entorno de las matemáticas.

Uno de los hechos que nos llevaron a tomar esta decisión ha sido el de poder contar desde el principio con la base de conocimientos donde se describen claramente los conceptos matemáticos con los que trabajamos. Las definiciones de estos conceptos nos han permitido implementar una semántica robusta, siendo ésta uno de los pilares fundamentales para el buen funcionamiento del sistema, tanto desde la perspectiva del módulo de deducción como desde la de la interfaz. Una vez disponible la semántica, junto con la jerarquía de los tipos, es sencillo escribir las entradas del diccionario PRÓGENES en el marco de las gramáticas categoriales de unificación.

Los buenos resultados conseguidos nos hacen pensar que este tipo de aproximación es muy adecuada para abordar problemas de procesamiento de lenguaje natural en los que el objetivo final sea principalmente la obtención de expresiones formales de cualquier tipo,

asociadas a la semántica de la oración. En este sentido resulta una aproximación muy conveniente en nuestro sistema, en el que del texto del problema se extrae la interpretación formal equivalente.

En nuestro caso, suponemos una interpretación única por construcción (los enunciados de los problemas no pueden ser ambiguos...), pero en otros entornos no sería problemático considerar que los textos a analizar tuvieran más de una posible formalización.

Evidentemente, todas las aproximaciones lexicalistas hacen que el mayor coste del sistema recaiga en el léxico, tanto en el sentido de asociar una gran cantidad de información a las palabras, como en garantizar una adecuada definición de las mismas. La precisión de las definiciones, además, se complica en lo relativo a la determinación de la semántica asociada a cada entrada. Por lo tanto, en este tipo de aplicaciones parece necesaria la participación de expertos en la materia que se quiere tratar, es decir, personajes capaces de determinar los conceptos fundamentales: su tipo y sus descriptores. En este sentido, el desarrollo de la interfaz de PRÓGENES involucra no sólo a las personas dedicadas a la propia interfaz, sino también a los matemáticos que llevan a cabo la determinación de los conceptos asociados al dominio.

El sistema se implementa sobre Risc System/6000, y el lenguaje de programación utilizado es el Common LISP de Lucid.

Referencias

[BIH88] Block, H.U., Haugenader, H.: "An Efficiency-oriented LFG PARSER". En: *Readings in NLP*; Ed.: B.J. Grosz, K. Sparck Jones & B.L. Webber, Morgan Kaufmann Publishers, Inc., 1986.

[CKZ88] Calder, J., Klein, E. y Zeevat, H.: "Unification Categorial Grammar: A concise, Extendable Grammar for Natural Language Processing". COLING 88, Budapest.

[KAY86] Kay, M.: "Parsing in functional unification grammar". En: *Readings in NLP;* Ed.: B.J. Grosz, K. Sparck Jones & B.L. Webber, Morgan Kaufmann Publishers, Inc., 1986.

[MIS91] Moreno-Torres, I. y Solias, M.T.: "Un analizador con 'chart' para gramáticas categoriales de unificación." En: *Procesamiento del Lenguaje Natural*, Boletín nº 9, págs.207-225. Ed: Sociedad española para el procesamiento del Lenguaje Natural, 1991.

[MCS91] Moriyón R., Castells P. y Saiz P.: "Base de Conocimientos PR_GENES." Ed: Informe interno IIC, 1991.

[REY88] Reyle, U.: "Compositional Semantics for LFG". En: *Natural Language Parsing and Linguistic Theories*, págs. 448-474, Ed.: D. Reidel Publishing Company, 1988.

[SHI86] Shieber, S.M.: "An Introduction to Unification-Based approaches to Grammar", Ed.: CSLI, Ventura Hall. Stanford University, 1986

[WIW90] Wittenburg, K. y Wall, R.E.: "Parsing with Categorial Grammar in Predictive Normal Form". En: *Inheritance hierarchies in Knowledge representation and programming languages*. Ed.: Wiley, 90.

[YIK90] Yampol, T. y Karttunen, L.: "An Efficient Implementation of PATR for Categorial Unification Grammar". En: *COLING 90* Vol 2. Ed.: Hans Karlgren.

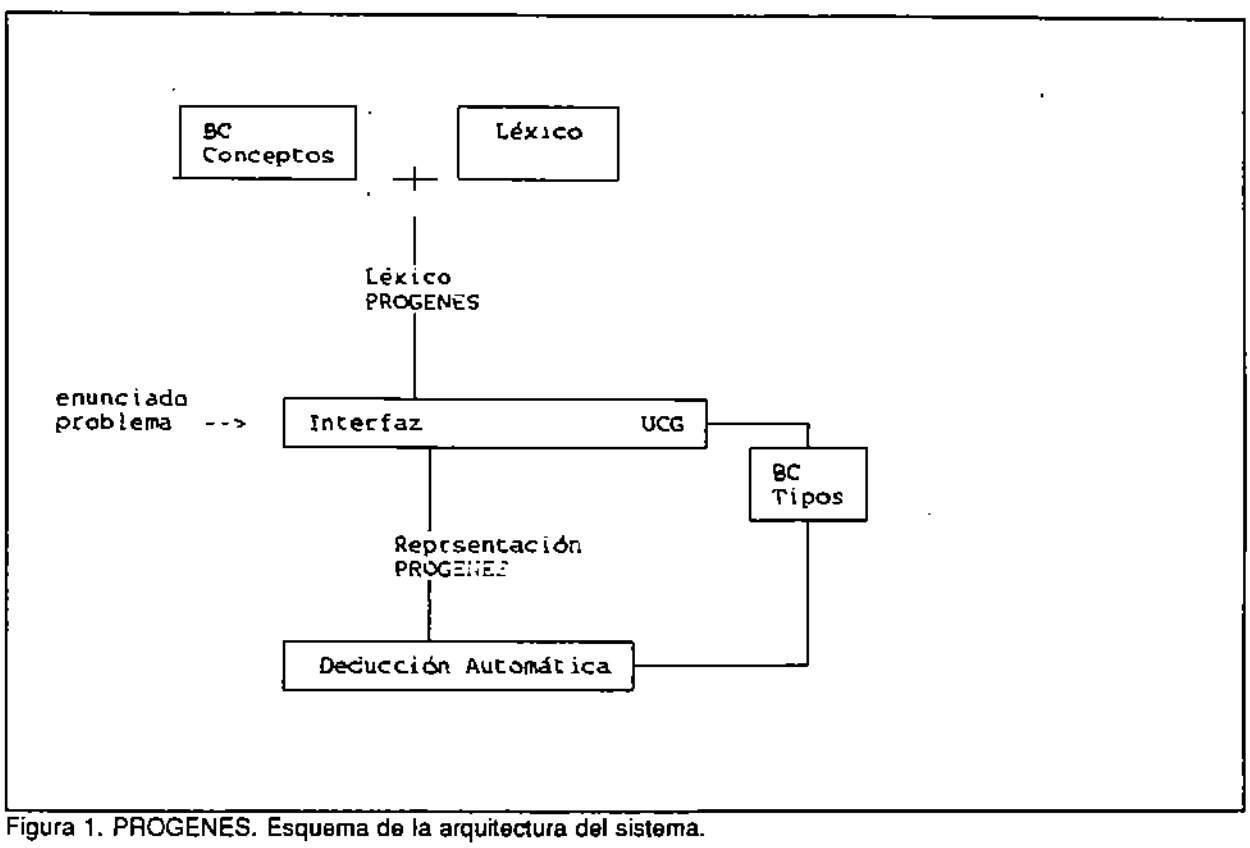

Figura 1. PROGENES. Esquema de la arquitectura del sistema.

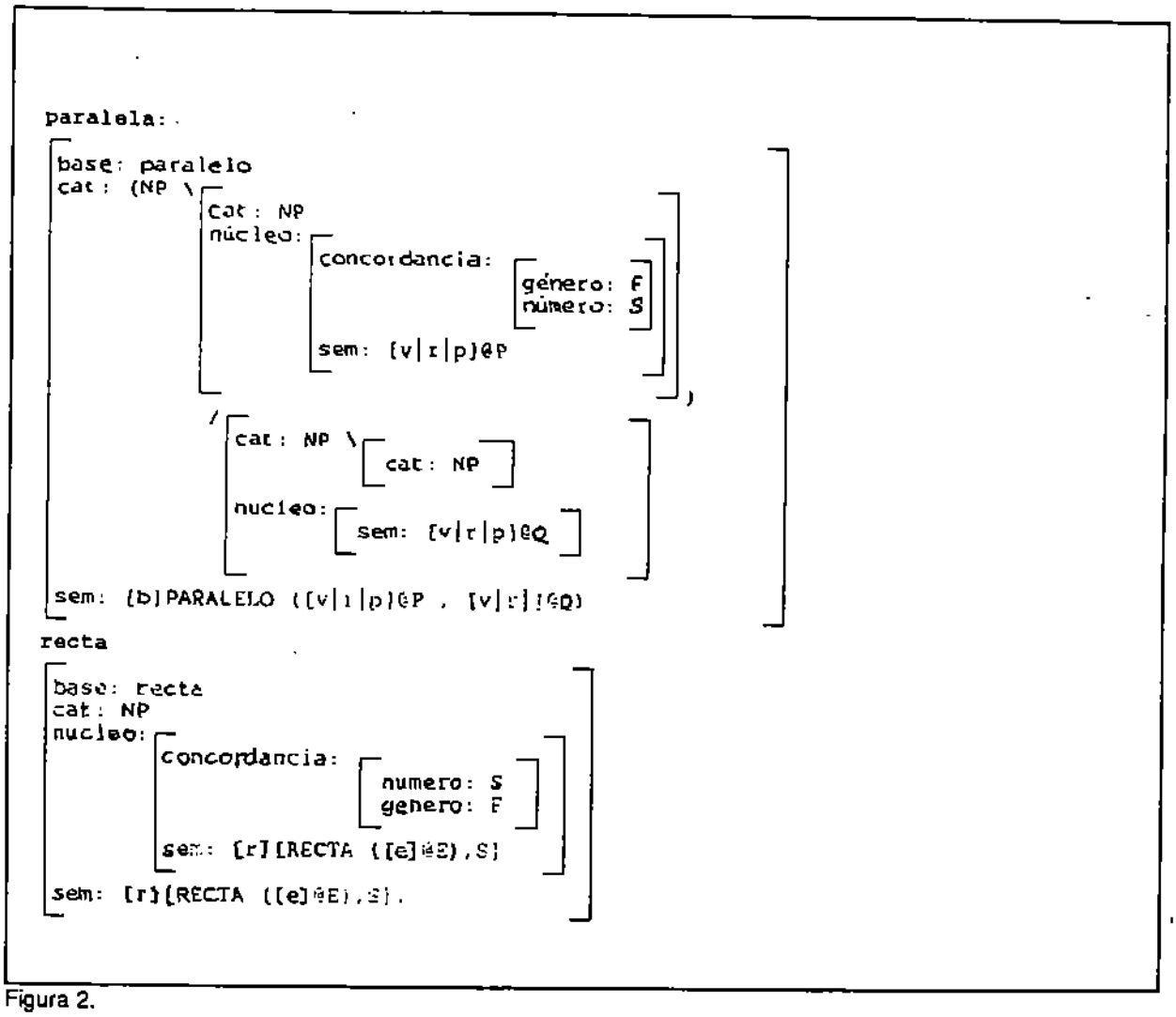
paralela:
base: paralelo
cat: (NP \
cat: NP
núcleo:
concordancia:
género: F
número: S
sem: [v|r|p]@P
)
/
cat: NP \
cat: NP
nucleo:
sem: [v|r|p]@Q
sem: (b)PARALELO ([v|r|p]@P , [v|r|p]@Q)
recta
base: recta
cat: NP
nucleo:
concordancia:
numero: S
genero: F
sem: [r][RECTA ([e]@E),S]
sem: [r][RECTA ([e]@E),S].

Figura 2.

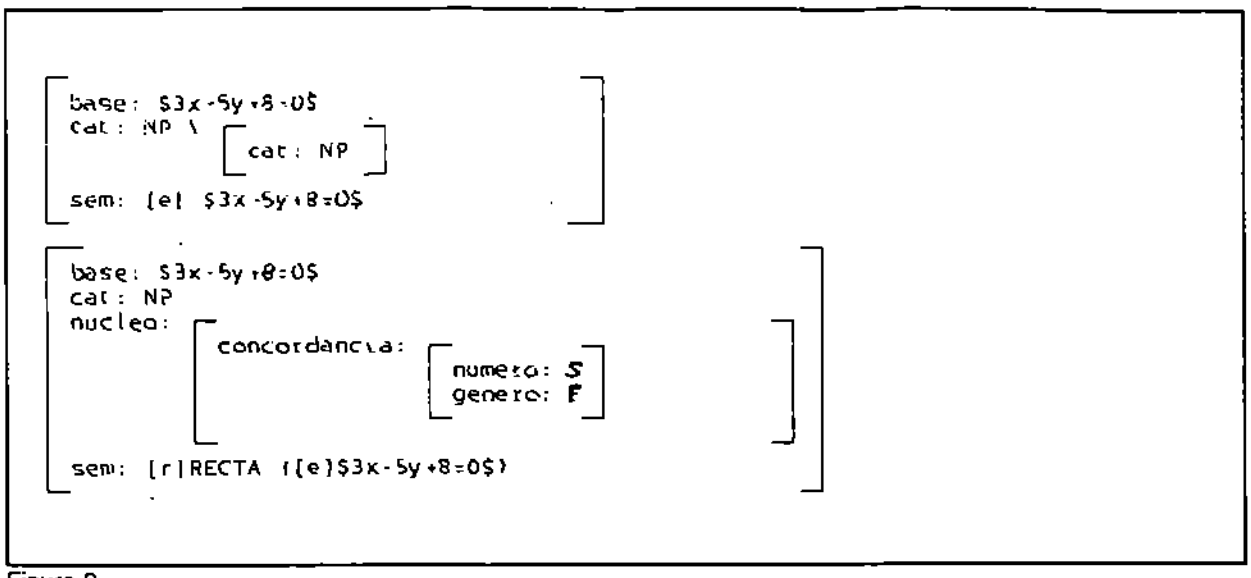

Figura 3.

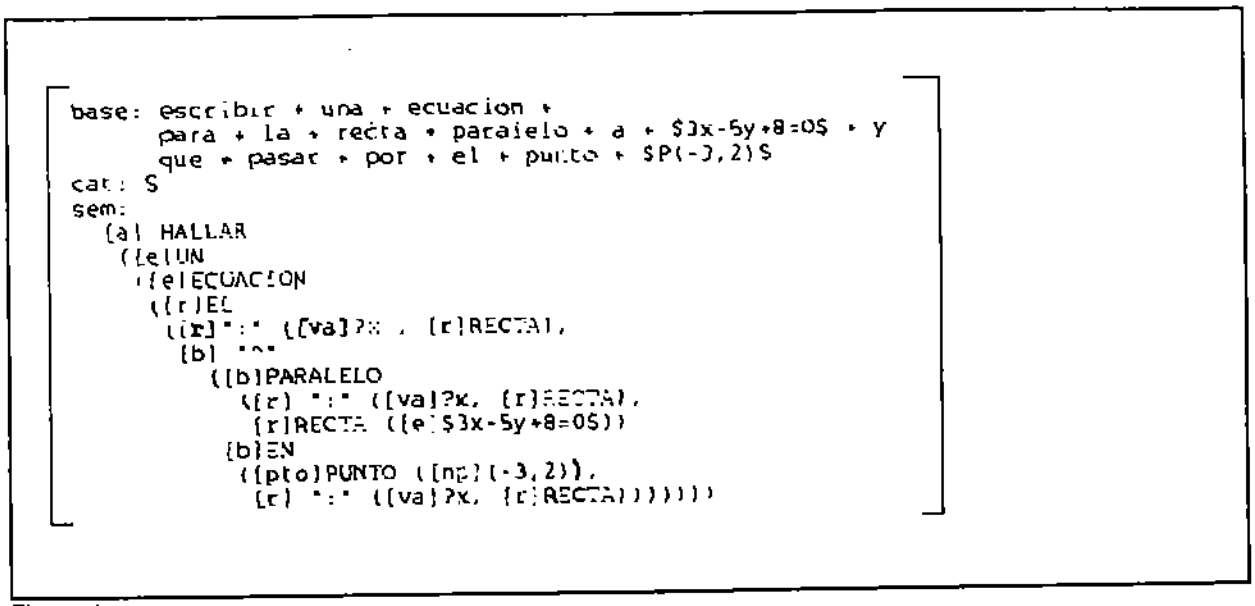

Figura 4.

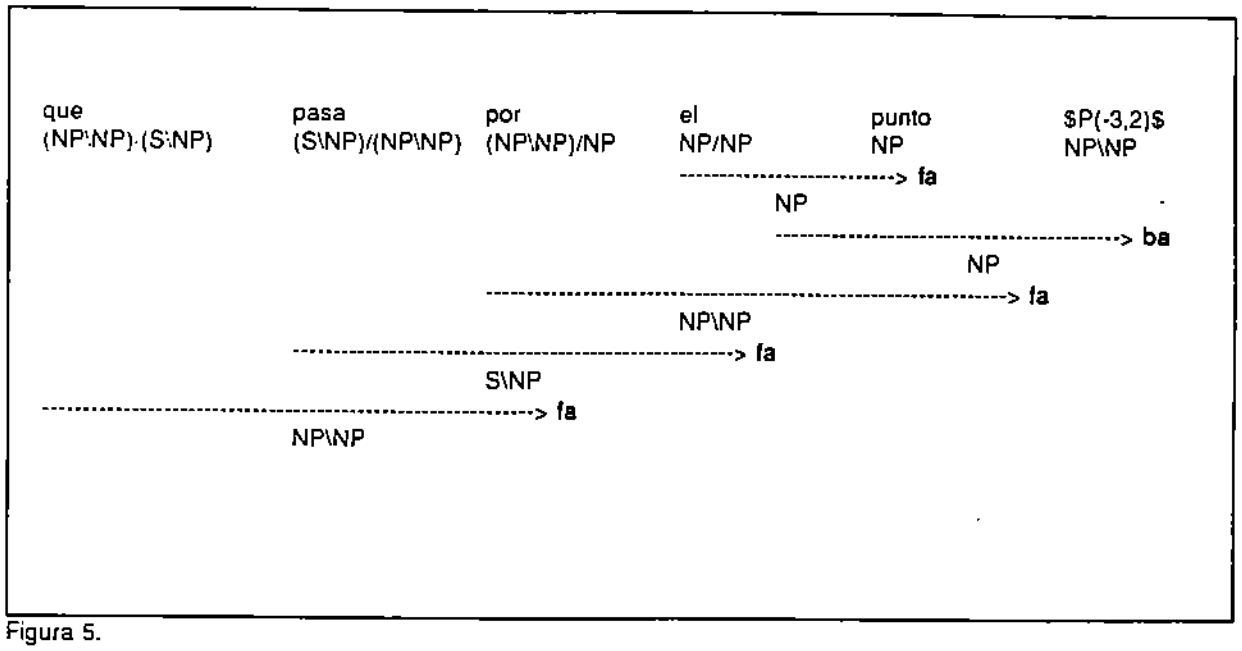
que
(NP\NP)/(S\NP)
pasa
(S\NP)/(NP\NP)
por
(NP\NP)/NP
el
NP/NP
punto
NP
$P(-3,2)$
NP\NP
fa
NP
ba
NP
fa
NP\NP
fa
S\NP
fa
NP\NP

Figura 5.