Revista de Filología y Linguística: XV (1) 99-105, 1989

GRAMATICAS INDIZADAS Y LENGUAS NATURALES (*)

Celso Vargas
Sección de Filosofía
Sede Regional de Occidente

ABSTRACT

In this paper we describe, in an informal way, the general form that takes the class of grammars known as indexed grammars. Given the correspondence between indexed grammars and the class of acceptors known as non-deterministic finite state machines with a pushdown store, these grammars constitute a formalism very fruitful for expressing linguistic theories.

1. LA JERARQUIA DE CHOMSKY.

A partir de 1957, fecha en que se inicia lo que ha sido denominado por muchos "la revolución chomskiana", el estudio de las propiedades matemáticas de los modelos lingüísticos que se proponen, y la caracterización matemática de la clase de posibles gramáticas, adquieren cada vez más importancia. Es deseable que una teoría sintáctica tenga una estructura matemáticamente precisa por razones como las siguientes: después de Chomsky se acepta que una lengua natural es un sistema caracterizado parcialmente por un conjunto de reglas que el niño trae de manera innata o que construye a partir de la experiencia tomando en cuenta su dotación genética.

Chomsky fue el primero que, sistemáticamente, señaló que las gramáticas pueden ser clasificadas en una escala de complejidad decreciente dependiendo del tipo de reglas que posea y la clase de lenguajes o lenguas que genere. Se conoce con el nombre de *Jerarquía de Chomsky* a tal clasificación. La clase de gramáticas más potentes que se conoce es la de las gramáticas tipo-0 y que generan lenguajes recursivamente enumerables y lenguajes recursivos. Las gramáticas tipo-1 o gramáticas dependientes de contexto, generan lenguajes dependientes de contexto; las gramáticas tipo-2 o gramáticas libres de contexto generan lenguajes libres de contexto; y, finalmente, las gramáticas tipo-3 o gramáticas regulares que generan lenguajes regulares. Esta jerarquía puede correlacionarse de manera consistente con las clases de procedimientos algorítmicos conocidos. Es decir, cada una de las clases de gramáticas tiene su correspondencia en términos de clases de procedimientos algorítmicos. Las gramáticas tipo-0 requieren todo el poder de las máquinas de Turing. Sin embargo, existe una diferencia importante entre las lenguas recursivamente enumerables y los lenguajes recursivos. En efecto, los lenguajes recursivamente enumerables no siempre son decidibles. En efecto, dada una secuencia determinada de símbolos (u oración) no siempre es posible decidir si pertenece o no al lenguaje generado por una gramática tipo-0. Un lenguaje es recursivo si y solo si es caracterizable en términos de un subconjunto propio de máquinas de Turing, esto es, para cada lenguaje recursivo existen dos Máquinas de Turing tales que una acepta dicho lenguaje y la otra el complemento; las dependientes de contexto por un subconjunto de máquinas de Turing deterministas pero de una complejidad menor que las requeridas por las clases anterior (esta complejidad viene determinada por el tiempo o espacio requerido para 'correr' el algoritmo). Existe una equivalencia entre estas Máquinas de Turing y las Máquinas de estado finito con dos pilas de memoria (two pushdown stores). Las gramáticas libres de contexto son equivalente a una máquina de estado finito determinista con una pila de memoria. Finalmente, las gramáticas regulares son enteramente caracterizables en términos de máquinas de estado finito sin memoria (véase Manna 1974 cap.1).

Tanto las gramáticas tipo-0 o recursivas como las gramáticas dependientes de contexto tienen un poder expresivo bastante grande que las hace candidatos "deficientes" para el análisis de las lenguas naturales. Por otro lado, las investigaciones empíricas recientes han mostrado o por lo menos sugieren fuertemente que las lenguas naturales no son libres de contexto (Véase Bresnan et.al 1982; Gazdar 1985).

En este contexto quisiera presentar de manera bastante informal una clase de gramáticas, conocidas como *Gramáticas Indizadas* que prometen bastante en la investigación sintáctica, aun cuando la tendencia en este momento, como señala Karttunen, "sugieren una conceptualización radicalmente diferente de cómo caracterizar las oraciones bien-formadas. La idea clave es la noción de información parcial y principios y restricciones mutuamente independientes" (Karttunen 1986: 9) y utilizando reglas exclusivamente libres de contexto. Sin embargo, esto no significa que esta clase no sea apropiada en este sentido, solo que esta nueva orientación puede requerir menos poder expresivo. Sin embargo, Shieber (1986) ha señalado que formalismos de este tipo pueden ser extendidos de tal manera que recorran toda la jerarquía de Chomsky.

Ofrezco primero una caracterización de esta clase de gramáticas y luego intentaré aplicar una de estas gramáticas en el tratamiento de algunos fenómenos lingüísticos.

2. LAS GRAMATICAS INDIZADAS.

El primero, al parecer en estudiar las gramáticas indizadas y a quien se debe el nombre es Aho (Aho 1968, citado en Gazdar 1985). Durante la presente década han sido ampliamente utilizadas en la descripción de varios fenómenos lingüísticos para los cuales las gramáticas libres de contexto fallan. Obtenemos una gramática indizada, de acuerdo con Gazdar si modificamos la clase de gramáticas libres de contexto de tal manera que "(i) se permita que las gramáticas empleen un rasgo distintivo único (single designed feature) tomado de una pila (stack) de ítemes a partir de un conjunto finito y que constituye sus valores; (ii) las reglas pueden 'empujar' ítemes dentro de la pila, sacar ítemes de la pila y copiar de la pila" (Gazdar 1985:1).

Las gramáticas que cumplen estas condiciones se ubican, en la jerarquía de Chomsky, entre las gramáticas libres de contexto y las gramáticas dependientes de contexto. Toda gramática libre de contexto es una gramática indizada, pero no al revés; pero además toda gramática indizada es una gramática dependiente de contexto pero la conversa no es válida.

Una gramática se define formalmente como un cuádruple <VN, Vt, S, R>, donde Vn es el vocabulario no terminal (en esta presentación utilizaremos A, B, C,... como vocabulario no terminal), Vt es el vocabulario terminal (utilizaremos a,b,c,...), S es axioma o punto de partida (en ocasiones se utiliza S' también (en español 0) y R es un conjunto de reglas de una clase específica. Una gramática indizada, en la definición estandar presentada por Gazdar hace uso además de las letras i,j,k como letras distinguidas que designan índices, las pilas y los períodos) [], [..], [i..]), donde [i..] es la pila cuyo índice superior es i; [] designa la pila vacía y [..] copia la pila tal y como está. Sea A un símbolo no terminal y W una variable que denota secuencias de símbolos terminales o no terminales adecuadamente especificados. Entonces, una gramática indizada tiene tres tipos básicos de reglas:

1- A [..] —> W [i..]
2- B.[i..] —> W [..]
3- C.[..] —> W [..].

La regla del primer tipo introduce en la pila un nuevo elemento; la segunda regla elimina de la pila el elemento que está más arriba y la regla tercera copia lo que hay en la pila rotulada como A a la pila rotulada como W. De acuerdo con Gazdar,

> Los índices que completan las pilas son extraídos de un vocabulario finito aunque las pilas mismas no tienen límite en tamaño. Cualquier límite superior en el tamaño de las pilas restringe la clase de gramáticas a la clase libre de contexto (Gazdar 1985: 3).

La clase de gramáticas caracterizables de este modo pueden correlacionarse de manera directa, esto es, pueden ser aceptadas por una máquina de estado finito no determinista con una pila de memoria. De hecho la clase de las gramáticas indizadas es capaz de generar todas las gramáticas que se ubican bajo la jerarquía. Esto es, toda gramática regular (sea de ramificación de derecha o de izquierda) puede ser generada por una gramática indizada, lo mismo puede decirse de las gramáticas libres de contexto y cierto subconjunto de las gramáticas dependientes de contexto.

Una máquina o autómata, en el sentido que aquí se utiliza, es un modelo matemático apropiado en la realización de ciertas tareas computacionales. Por ejemplo, dada una expresión cualquiera, existe un autómata que permite decidir si esta expresión puede ser generada por una gramática libre de contexto o no, dada la equivalencia entre gramáticas libres de contexto y cierta clase de autómatas. Una máquina está constituida por: (1) un conjunto de estado S tales que existe un estado inicial (punto de partida) y uno o más estados finales (el estado inicial puede ser tanto inicial como final) y un conjunto de estados intermedios; (2) una función f que especifica como pasar de un estado a otro; (3) un alfabeto cuyos elementos aparecen en la expresión a considerar, es decir, constituyen el 'input' de la máquina. Se dice

que una máquina es no determinista si para pasar del estado unicial al estado final de la máquina existe un conjunto de vías (paths). Es decir, la máquina puede utilizar cualquiera de estas vías para llegar a la solución del problema. La máquina puede seleccionar arbitrariamente cualquiera de estas vías.

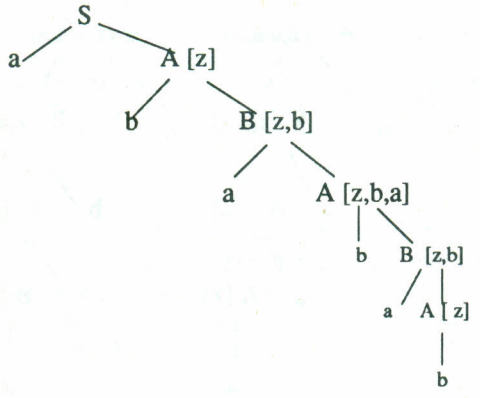
Quisiéramos presentar ahora algunos ejemplos de gramáticas que pueden ser expresadas dentro de esta clase. Si Vt = a, b , entonces, un lenguaje es cualquier subconjunto formado por secuencias a partir de Vt. Por ejemplo, el conjunto a, abab, abababab,... es un lenguaje. Podemos expresar este lenguaje como (ab)* (abreviación utilizada por Manna (1974) para referirse a lenguajes regulares). Este lenguaje puede ser generado por la siguiente gramática indizada:

4- a. S —> a A [z,..]
b. A [..]—> b B [b,..]
c. B [b,..]—> a A [..]
d. B [..] —> a A [a,..]
e. A [a,..] —> b B [..]
f. A [z,..] —> b.

En este ejemplo, 'z' es el parámetro para determinar cuando debe detenerse la recursividad. Es decir, cuando el último elemento de la pila es z, entonces, por regla e es sacado y es reemplazado por una b. La secuencia 'ab'es generada del siguiente modo: aplicamos primero regla a y guardamos en la pila el elemento z, aplicamos luego la regla f que saca z y lo reemplaza por b. Podemos expresar en forma gráfica el lenguaje generado por esta gramática mediante la derivación de la secuencia 'ababad':

Este mismo lenguaje puede ser generado por una gramática regular como la siguiente:

5.



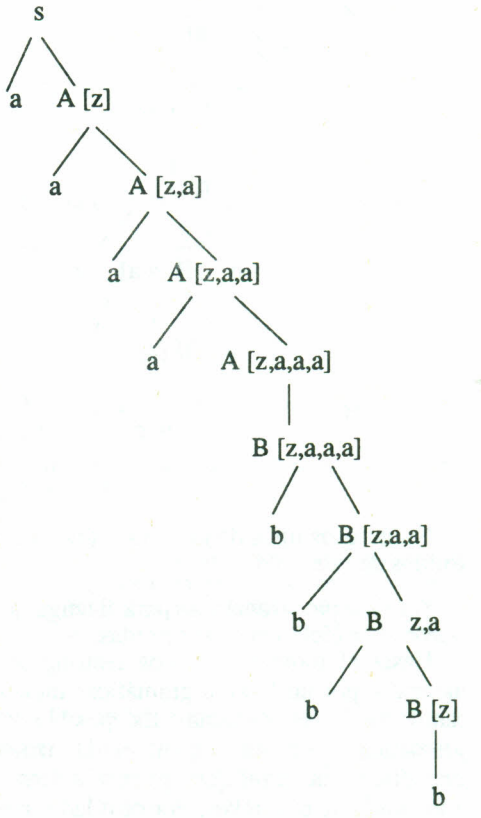
6. a. S —> a A
b. A —> b B
c. B —> a A
d. A —> b.

Como se puede ver 6 es más simple que 4. Sin embargo, 4 es más interesante en tanto que nos ilustra como utilizar las pilas de memoria. La siguiente gramática genera un lenguaje a^nb^n que es un lenguaje libre de contexto:

7. a. S —> a A [z,..]
b. A [..]—> a A [a,..]
c. A [..] —> B [..]
d. B [a,..] —> b B [..]
e. B [z,..] —> b

Esta gramática genera árboles como el siguiente:

8.



Resulta mucho más fácil formular una gramática que genere ese lenguaje en términos libres de contexto. De hecho requiere solo dos reglas:

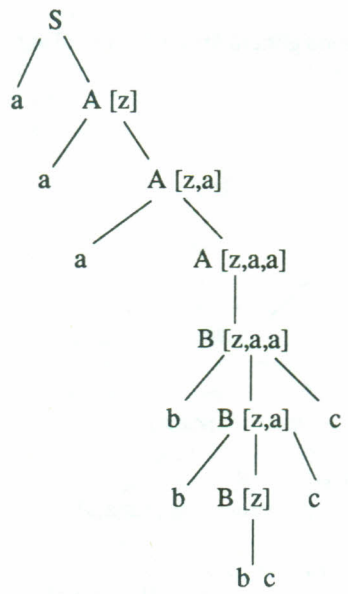
9. a. S —> a S b
   b. S —> a b.

La siguiente gramática genera un lenguaje de la forma $a^n b^n c^n$ que es un lenguaje dependiente de contexto:

10. a. S —> a A [z,..]
    b. A [..] —> a A [a,..]
    c. A [..] —> B [..]
    d. B [a,..] —> b B [..]c
    e. B [z,..] —> b c

Dicha gramática genera árboles como el siguiente:

11.



Numerosos tipos de gramática pueden ser construidos de este modo.

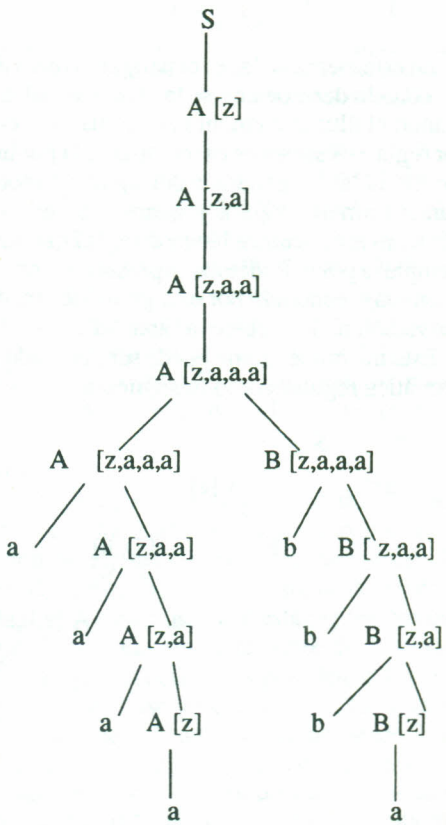
Por ejemplo, gramáticas para el lenguaje $a^n b^n c^n d^n$ pueden ser fácilmente construidas.

Hasta el momento hemos restringido nuestra atención, por un lado, a gramáticas indizadas que ramifican a la derecha. Es posible construir gramáticas que ramifiquen a la izquierda o gramáticas que ramifiquen parte a la derecha, parte a la izquierda o al revés. Por otro lado, nos hemos limitado a las gramáticas en las que aparece un único símbolo no terminal en cada regla (A o B). De hecho no hay necesidad de restringirnos en ese sentido. Una gramática en la que aparece más de un símbolo no terminal en algunas reglas aumenta la complejidad de la gramática pero no la clase de lenguajes generados (al menos eso parece. Véase Gazdar 1985 apéndice). Por ejemplo, el lenguaje $a^n b^n$ descrito anteriormente también puede ser descrito del siguiente modo:

12. a. S —> A [z,..]
    b. A [..] —> A [a,..]
    c. A [..] —> A [..] B [..]
    d. A [a,..] —> a A [..] b B [..]
    e. A [z,..] —> a
    f. B [z,..] —> b

Una gramática así formulada es relativamente compleja cuando se expresa en forma de árboles derivacionales. En efecto, la secuencia aaaabbbb generada por esta gramática tiene la siguiente estructurada:

13.



Es posible esperar que para cada gramática indizada en la que aparece más de un símbolo no terminal por regla, podamos construir la correspondiente

gramática indizada que hace uso de un único símbolo no terminal por regla (véase Gazdar 1985 apéndice).

3. APLICACIONES DE LAS GRAMATICAS INDIZADAS.

Las gramáticas indizadas han tenido muchas aplicaciones en el tratamiento de algunos fenómenos lingüísticos (véase Gazdar 1985; Gazdar y Pullum 1985 para referencias). Intentaremos en este apartado aplicar una gramática indizada en el tratamiento de la serie de dependencias cruzadas del holandés. Los datos que se van a utilizar los tomo de Bresnan et al. (1982). En ese trabajo Bresman et al. se plantean dos metas fundamentalmente: por un lado, mostrar que las series de dependencias del holandés no pueden ser fuertemente generados por una gramática libre de contexto, es decir, una gramática libre de contexto está imposibilitada de asignar las descripciones estructurales correctas a esas dependencias. En segundo lugar, muestran que dichas dependencias pueden ser fuertemente generadas por una gramática léxico-funcional (GLF). El estudio de las dependencias cruzadas de lenguas como el holandés se ha constituido en uno de los 'procedimientos de prueba'de las teorías sintácticas. De ahí el interés de este 'fenómeno' para la presente discusión.

Las GLFs utilizan dos tipos de reglas: el primero de ellos son reglas libres de contexto que generan las estructuras-c (estructuras de constituyentes) de las oraciones. Existe un segundo tipo de reglas que toman estas estructuras y las proyectan en un análisis funcional (estructuras-f).

Las estructuras-f son estructuras jerárquicas que representan formalmente las relaciones gramaticales de las oraciones en términos de funciones universales como SUB (eto), OB (jeto) y COMP (lemento) abstrayendo las diferencias de las lenguas particulares en la estructura superficial (Bresnan et. al 1982: 624).

Para que una secuencia sea gramatical deben cumplirse dos condiciones: por un lado, tiene que ser generada por la gramática y, por el otro, las estructuras-f asociada con esa secuencia tiene que cumplir las condiciones de Unicidad, Completitud y Coherencia.

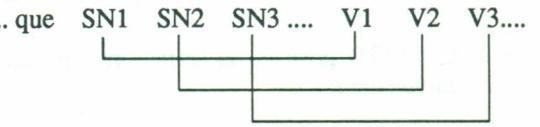
De acuerdo con varios comentarios hechos por Perrault (1985) las GLFs tienen un poder expresivo que va más allá de las gramáticas indizadas. En efecto:

[...] Berwick y Roach han examinado la relación entre GLF y la clase de lenguajes generados por las *gramáticas indizadas*, una clase que se sabe es un subconjunto propio de los LDCs (lenguajes dependientes de contexto), pero que incluye algunos lenguajes completos-PN. Afirman (comunicación personal) que los lenguajes indizados son un subconjunto propio de las GLFs (Perrault 1985: 14).

Los lenguajes completos-PN son aquellos que pueden ser reconocidos por una Máquina de Turing con un tiempo específico y que requieren que se tome en cuenta todas las posibilidades. Sin embargo, aquí no nos ocuparemos de las GLFs, sino que utilizaremos una gramática indizada para dar cuenta de estas series de dependencias.

El holandés exhibe series de dependencias cruzadas entre SN y V cuando aparecen en oraciones incrustadas. La estructura es la siguiente:

14.



un ejemplo en cuestión es el siguiente:

15.

'...que Jan Marie los niños Piet ver-pas ayudar hacer nadar'

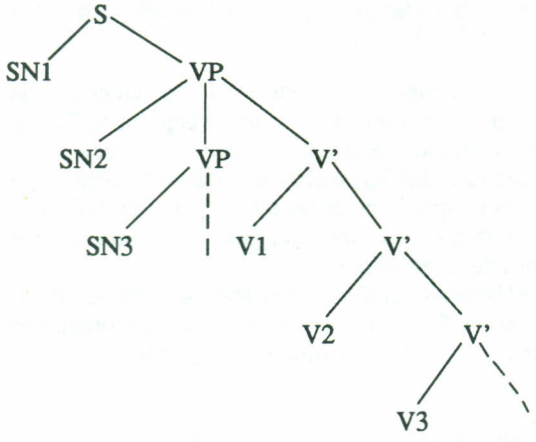
'...que Jan vio a María ayudar a los niños para hacer que Piet nade'

El verbo en primera posición es formalmente distinguido porque marca el tiempo y la persona y la concordancia en número con el primer SN. El verbo en la última posición es distinguido de los otros mediante restricciones de subcategorización (Bresnan et. al 1982: 615).

De acuerdo con los autores, si mantenemos las restricciones indicadas, esto es, la concordancia entre el primer SN y el primer verbo de la serie, entre el último SN y el último verbo de la secuencia, podemos intercambiar de posición a los verbos restantes y todas las oraciones resultantes son gramaticales (véase p.615). Esto significa que una gramática indizada para esas series debe respetar y expresar

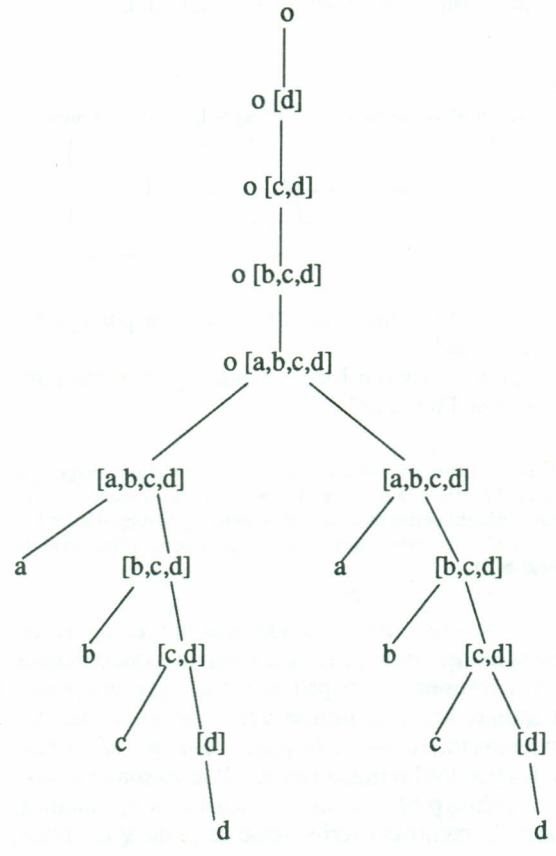
estas restricciones. Los autores, afirman que la estructura sintáctica correcta de estas series es:

16.



Gazdar (1985) presenta la estructura siguiente como equivalente a 16.

17.



Suponiendo que esta equivalencia es correcta, propongo la siguiente gramática indizada que daría cuenta de las series de dependencias cruzadas del holandés:

18. a. S —> S [1,..]

b. S [..] —> [i + 1,..]

c. S [i,..] —> SN [i..] SV [i,..]

d. SN [i,..] —> SN SN [..]

$$\begin{bmatrix} \alpha\ pl \\ \beta\ pers \\ \gamma\ gen \end{bmatrix}$$

e. SV [i,..] —> V V [..]

$$\begin{bmatrix} \alpha\ pl \\ \beta\ pers \\ \gamma\ gen \end{bmatrix}$$

f. SN [1,..] —> SN

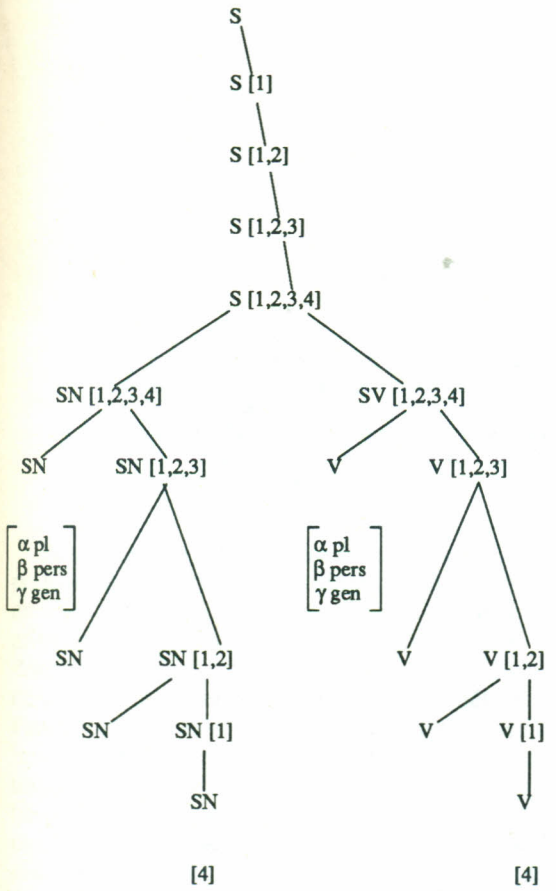
[i]

g. V [1,..] —> V

[i]

En la formulación de estas reglas he utilizado α para + - plural, ß para primera, segunda o tercera persona y γ para marcar el género (masculino, femenino, neutro). Se utiliza 'i' como un índice numérico que sirve como contador. De acuerdo con este fragmento de gramática las dependencias exhiben la siguiente estructura:

19.



La clase de las gramáticas indizadas es sumamente importante para las lenguas naturales. Según parece no se conoce ningún caso, hasta el momento, que no pueda ser tratado con el poder expresivo de alguna gramática indizada. Las gramáticas indizadas poseen, desde el punto de vista de su estructura matemática, una complejidad manejable desde el punto de vista algorítmico. Conforme se recorre la jerarquía hacia arriba la complejidad de las gramáticas aumenta y con ello surgen dificultades en su manejo algorítmico. Quizá deba esperarse que los distintos modelos lingüísticos que se presenten deban ajustarse a la clase de gramáticas indizadas. Esto puede hacerse de varias maneras: por un lado, formulándolos dentro de una gramática indizada, o por otro lado, mostrando que la formalización de ese modelo puede expresarse en términos de una gramática indizada, es decir, para cada modelo lingüístico propuesto, existe una gramática indizada equivalente. O, finalmente, para cada modelo propuesto, existe una máquina de éstado finito no determinista que acepta todas las secuencias generadas por ese modelo. Pero dada la equivalencia entre este tipo de máquinas y las gramáticas indizadas, debe existir un procedimiento para hacer que esa gramática sea formulable en términos de una indizada.

(*) Agradezco a Edwin Bonilla la lectura a un borrador anterior y sus valiosos comentarios.

BIBLIOGRAFIA

Bresnan, Kaplan, Peters and Zaenen (1982) "Cross-serial Dependencies in Dutch" en: *Linguistic Inquiry* (13) 613-635.

Gazdar, Gerard (1985) *"Applicability of indexed grammar to natural languages." CSLI, Stanford University.*

_____ y Pullum (1985) *"Computationally Relevant Properties of Natural Languages and their Grammars." CSLI, Stanford University.*

Karttunen, Lauri (1986) *"The Relevance of Computational linguistics." KSLI, Stanford University.*

Manna, Zohar (1974) *Mathematical Theory of Computation.* McGraw-Hill, Inc. USA.

Perrault, Raymond (1984) *"On the Mathematical Properties of Linguistic Theories." CSLI, Stanford University.*