Sobre la naturaleza de la información lingüística y la relación entre las gramáticas categoriales y las gramáticas independientes de contexto

J.A. Jiménez Millán. Universidad de Cádiz.
Escuela U. Politécnica de Cádiz. Calle Chile s/n. 11003 - Cádiz.
Tfno:(956)221960 Fax:(956)224359 email: jose@hercules.uca.es

J.M. Guirao Miras. Universidad de Granada.
Escuela T. Superior de Ingeniería Informática.
Tfno:(958)242813 email:jmguirao@ugr.es

8 de junio de 1995

Resumen

En este trabajo presentamos una extensión de la teoría de las *Gramáticas Categoriales* (GC) que se puede resumir en el slogan: *Gramáticas Categoriales + relación de orden parcial* ($GC_{\leq}$). Pretendemos que esta extensión popularice su utilización.

También analizamos la relación entre las $GC_{\leq}$ y cierto subconjunto de las gramáticas independientes de contexto (GIC), para lo cual hemos desarrollado un modelo algebraico común a ambas teorías (con una posible ampliación a las gramáticas sensibles al contexto GSC). Con esto pretendemos constatar la equivalencia de ambos formalismos en su poder generativo fuerte.

El modelo puede entenderse como una fundamentación algebraica de las gramáticas categoriales con una clara inspiración intuitiva; de forma que las reglas de inferencia de las GC se pueden deducir de la construcción matemática del modelo.

Una característica importante de la teoría es que durante el desarrollo se separan las propiedades que dependen, exclusivamente, de la naturaleza de la información lingüística de aquellas propiedades que dependen de características específicas de cada lenguaje.

Palabras clave: Gramáticas categoriales. Modelos matemáticos de gramáticas. Gramáticas independientes de contexto. Teoría de la información. Dominios de Scott.

1 Introducción

Presentamos, en este artículo, una extensión de la teoría de las gramáticas categoriales (GC) que ha surgido al investigar la relación entre las GC y las gramáticas independientes de contexto (GIC) y, en concreto, al intentar establecer un algoritmo automático para pasar las descripciones gramaticales de uno a otro formalismo.

Durante la investigación encontramos que las GIC son un formalismo más potente que las GC debido a dos razones:

- Las GIC no son capaces de discriminar entre gramáticas significativas y gramáticas no significativas (desde un punto de vista lingüístico). Esta no es una característica deseable a un formalismo.

- Existe un fenómeno de *jerarquía* en las GIC relacionado con la *herencia* entre categorías sintácticas y con la naturaleza *incompleta* de la información lingüística. Fenómeno que no aparece en las GC. Esto hace que las GIC sean más intuitivas y fáciles de utilizar que las GC. Como resultado proponemos una extensión a las GC en la que se les dota de la relación jerárquica que les falta.

  Recordemos que, en el formalismo de las GC, la información gramatical está restringida al lexicón mientras que las reglas de deducción proporcionan, exclusivamente, la lógica de la operación de concatenación entre las categorías sintácticas, pero sin reflejar ningun tipo de conocimiento gramatical de ninguna lengua concreta. Proponemos incluir cierto conocimiento gramatical en forma de una relación de orden parcial. Esta información emergerá hasta las propias reglas de deducción a través de la maquinaria de subsunción de categorías.

  La relación de orden está definida mediante dos teorías diferentes: Una teoría de la lógica subyacente a la operación de concatenación de estructuras sintácticas, y otra teoría jerárquica que refleja conocimientos sintácticos específicos de la lengua concreta que estemos considerando.

Vamos a comenzar, pues, analizando la relación entre las GIC y las GC. Después construiremos un modelo algebraico para las GIC (con una posible ampliación a las GSC) para, a continuación, proponer la mencionada extensión para las GC ($GC_{\leq}$). Esta extensión surge al utilizar una especialización del modelo algebraico desarrollado para las GIC como modelo para las GC. Es decir, utilizamos un modelo algebraico común para la extensión $GC_{\leq}$ y para aquellas GIC que cumplan ciertas restricciones matemáticas. De esta manera se comprueba la equivalencia de ambos formalismos (en el poder generativo fuerte).

2 Sobre la relación entre las GC y las GIC

En la literatura podemos encontrar que se ha intentado establecer, repetidas veces, la relación entre las gramáticas independientes de contexto (GIC) y las gramáticas categoriales (GC). Esta empresa se remonta a los años 1960 y 1966 con la llamada "Conjetura de Chomsky" que afirma que el cálculo de Lambek reconoce, justamente, los lenguajes independientes de contexto [5]. Véase Bar–Hillel [3], Buszkowski [4] y su bibliografía, Van–Benthem [18], y más recientemente Pentus [13].

Por un lado las GIC son las más utilizadas en la práctica, tanto para describir lenguajes formales como para describir lenguajes naturales (están mejor estudiadas y sabemos cómo construirlas de manera que recojan las propiedades deseadas como asociatividad, prioridad etc.), mientras que las GC tienen un fuerte atractivo teórico que permite manejar sus propiedades de forma abstracta y realizar demostraciones basadas en razonamientos algebraicos. Descubrir cual es la relación en el poder expresivo de ambas gramáticas permitiría utilizar el formalismo descriptivo adecuado con conocimiento de causa.

Generalmente se acostumbra a señalar que las gramáticas categoriales tienen el mismo poder generativo débil que las gramáticas independientes de contexto. Esto quiere decir que ambas son capaces de generar las mismas cadenas superficiales.

El tema es complejo. Por ejemplo, en Bar–Hillel y colaboradores [3] se demuestra que el cálculo de Lambek monodireccional es equivalente, en capacidad generativa débil, a las GIC; pero en Van–Benthem [18] podemos encontrar que al añadir la regla de permutación al anterior cálculo de Lambek se pueden reconocer gramáticas que no son independientes de contexto.

No obstante no nos basta con la capacidad generativa débil, y querríamos conseguir que a la misma cadena se le asigne el "mismo significado". Uno de los problemas es que no está claro el sentido que tenga la palabra "significado". Es la idea subyacente al poder generativo fuerte pero este concepto solo está bien definido para gramáticas transformacionales y, aunque podría ampliarse para gramáticas del mismo tipo, se presentarían problemas en el caso de dos formalismos diferentes.

Nuestra idea es que, tal y como afirman Wilks [19] y muchos otros, el análisis sintáctico es, fundamentalmente, una cuestión de asignar estructuras semánticas a los textos. Si utilizamos el principio de composicionalidad semántica, las estructuras semánticas se obtienen a partir de las estructuras sintácticas. Es decir, en el caso de las gramáticas independientes de contexto, la semántica se obtiene a partir del árbol de análisis. En el caso de las gramáticas categoriales, en que se suele manejar simultaneamente una componente sintáctica y otra semántica (el etiquetado), esto es más que evidente.

Como no nos basta con tener la misma capacidad generativa, en sentido

débil, y queremos que también se mantenga la semántica (aunque esto no está bien definido para formalismos diferentes), proponemos otro concepto que esperamos esté estrechamente relacionado. Lo que necesitamos es un modelo único, y no trivial, valido tanto para las GIC como para las GC. Este modelo debe reflejar todas las propiedades de ambos formalismos.

Buscaremos un algoritmo para construir este modelo. Notar que estamos hablando de una relación entre formalismos, y no entre gramáticas.

3 La estructura de la información lingüística

Antes de construir el modelo vamos a analizar la estructura matemática de la información lingüística.

1. Consideremos el conjunto BASCAT de categorías sintácticas básicas. Este conjunto estaría formado, por ejemplo, por las categorías {O, SN, SV, SAdj, N, ..., $\top$, $\bot$, ξ }, correspondientes a oración, sintagma nominal, sintagma verbal, sintagma adjetival, nombre común ... respectivamente, y las categorías correspondientes a *top*, *bottom* y *cadena–vacía* que están explicadas más adelante.

2. A continuación consideremos el conjunto BASCAT$^+$, formado por la clausura inductiva del conjunto BASCAT bajo la operación de concatenación ($\cdot$).

   Pediremos que esta operación de concatenación ($\cdot$) sea asociativa:

$$(x \cdot y) \cdot z = x \cdot (y \cdot z)$$

Si estamos de acuerdo, con Pollard y Sag [15]; Shieber [17]; Daelemans, De Smedt y Gazdar [7], y muchos otros; en que la información lingüística es de naturaleza *incompleta*, que viene dada por *restricciones*, y que existe un fenómeno de categorización y subcategorización en el que se observa el efecto de *herencia*; entonces podríamos definir un orden parcial en este conjunto BASCAT$^+$ intentando capturar el sentido intuituivo de información más o menos definida.

3. Para establecer una relación de orden, comenzaremos definiendo un elemento inicial máximo llamado *top*, $\top \in$ BASCAT, que representaría la ausencia total de información lingüística (nuestra cadena puede pertenecer a cualquier categoría posible).

   Al ir obteniendo más información, mediante restricciones y otros mecanismos lingüísticos, iríamos restringiendo más y más las posibles categorías compatibles con nuestro sistema. Así iríamos avanzando en sentido descendente en el espacio BASCAT$^+$.

4. También definimos un elemento terminal mínimo *bottom*, $\perp \in$ BASCAT, que se supone representa el exceso de información (la información contradictoria, no existe ninguna categoría compatible con ella).

   la categoría *bottom* se propaga respecto a la operación de concatenación:

   $\perp \cdot x = x \cdot \perp = \perp; \ \forall x \in \text{BASCAT}^+$

5. En el conjunto BASCAT$^+$ definimos una relación de orden parcial $x \geq y$ con el significado de información compatible y más o menos definida. En nuestro caso la categoría x sería compatible con la categoría y pero x más general que y (y más definida que x).

   Pediremos–por definición–que la relación de orden sea isótona respecto a la operación de concatenación. Es decir, que si se cumple que $y \geq z$, entonces $\forall x$ se debe cumplir que: $x \cdot y \geq x \cdot z$; y que: $y \cdot x \geq z \cdot x$.

6. La relación de orden parcial establecia por los puntos anteriores, es tal que nos permite definir la operación *ínfimo* ($\wedge$). Dos categorías cualesquiera x e y pertenecientes a BASCAT$^+$ siempre tienen un ínfimo, $x \wedge y$, aunque este puede ser *bottom*, $\perp$.

   Este elemento representa la categoría más general de las que son compatibles con la información suministrada por x e y.

7. Análogamente podemos definir, en el conjunto BASCAT$^+$, otra operación ($\vee$) llamada *supremo*. Dos categorías cualesquiera x e y siempre tienen un supremo, $x \vee y$, aunque este puede ser *top*, $\top$.

8. Podemos considerar una categoría especial llamada ξ (que no tiene representación externa), y que es el elemento neutro para la concatenación: $x \cdot \xi = \xi \cdot x = x$.

La relación de orden parcial establecida mediante los anteriores puntos (3) al (8), dotan al conjunto BASCAT$^+$ de una estructura de retículo. Nuestra particular fuente de inspiración en este caso han sido, entre otros, los trabajos de Hassan Aït-Kaci y, en concreto, los *sorts* de su lenguaje Life (véase por ejemplo [1]), así como los sistemas de información de Dana Scott [16].

A dicha relación de orden le llamaremos *orden de la lógica subyacente*, y solamente refleja conocimientos generales sobre la estructura incompleta de la información lingüística; pero no incluye ningún conocimiento específico de ninguna lengua concreta. Es necesaria información adicional que vendrá expresada mediante restricciones extras (una teoría que se construye sobre los objetos dotados ya de un orden subyacente).

La forma concreta que adopten las restricciones variará dependiendo de cada teoría gramatical. Así podrían adoptar la forma de ecuaciones,

inecuaciones, pertenencia a un conjunto... etc. Cada forma de expresar las restricciones tendrá su expresividad y su dificultad de implementación sobre un ordenador pero, por ahora, estos problemas no son relevantes.

9. Unas de las restricciones más importantes, y que llamaremos *restricciones jerárquicas*, tienen la forma de inecuaciones entre categorías del espacio BASCAT$^+$ — es decir, tienen la forma $x \geq y$, donde x, e y pertenecen a BASCAT$^+$ — intentando reflejar el fenómeno de la *jerarquía* y *herencia* en la información lingüística.

   No todas las restricciones que podamos inventarnos son lingüísticamente válidas.

   Si definimos un *semiretículo*$_\wedge$ como un conjunto ordenado que tiene definida la operación de ínfimo ($\wedge$) para cualesquiera dos elementos. Entonces, como una exigencia muy importante, pediremos que las restricciones jerárquicas sean compatibles con el orden de la lógica subyacente (es decir, que la operación de concatenación siga siendo isótona), y que el conjunto BASCAT$^+$ tenga la estructura de semiretículo$_\wedge$.

   En el capítulo siguiente ahondaremos en este tema.

   En el caso de que pidiésemos que la estructura resultante fuese también cerrada respecto a la operación de supremo $\vee$ (es decir, que adoptara la estructura de retículo), entonces conseguiríamos que la gramática no fuese ambigua (aunque el requisito primero es más fuerte que el último).

10. Definimos un *ideal*$_\wedge$, I, como un subconjunto $I \subseteq A$, tal que:

    a) si $z \in I$, entonces para $\forall k \in A$ tal que $k \leq y$ se debe cumplir necesariamente que $k \in I$.

    b) I es cerrado respecto a la operación de *ínfimo*, $\wedge$. Notar que no exigimos que sea cerrado respecto a la operación de *supremo*, $\vee$ (como es lo habitual).

    En nuestra estructura de retículo, toda categoría, x, genera un ideal$_\wedge$. Utilizaremos la notación $\Delta(\alpha)$ para referirnos al ideal$_\wedge$ generado por α.

    Cuando afirmamos que una cierta cadena de símbolos pertenece a la categoría sintáctica, x, en realidad queremos decir que pertenece al ideal$_\wedge$ generado por x, que está formado por todas las categorías sintácticas compatibles con ella.

    De esta manera definiremos el conjunto $\mathcal{CAT}$, de categorías sintácticas, como el conjunto formado por los ideales$_\wedge$ de BASCAT$^+$. El conjunto $\mathcal{CAT}$ es también un conjunto ordenado, pero bajo la relación de inclusión.

En lo sucesivo, cuando hablemos de una cierta categoría sintáctica, en realidad nos referiremos al conjunto CAT. No obstante, debido a que existe una correspondencia biunívoca entre los elementos de BASCAT$^+$ y sus ideales$_\wedge$ principales generados por estos elementos, a menudo abusaremos un poco del lenguaje hablando de una cierta categoría del conjunto BASCAT$^+$ cuando, en realidad, nos referimos al ideal$_\wedge$ principal generado por dicha categoría. Esperemos que este abuso del lenguaje no cause confusión y que su significado esté siempre claro por el contexto.

4 Lectura de una GIC como relación de orden parcial: Modelo algebraico para las GIC

Creemos que la dificultad que ha existido al intentar encontrar una relación entre las GIC y las GC se debe a la falta de formalización de las primeras. En efecto; es conocido que la notación de las GIC tiene su inspiración en la lógica y concretamente en la teoría de los sistemas combinatorios de Post (por ejemplo, Chomsky en [6](pp:8-9) hace referencia a ella, y Backus en [2] admite haberla utilizado), pero aun así su relación con la lógica no está clara. Pereira y Warren en [14] formalizan un sistema deductivo relacionando el análisis sintáctico con la demostración de un teorema y proporcionando una regla de inferencia (la deducción de Earley); pero no relacionan la gramática, en sí, con ningún sistema lógico conocido. Pertee y colaboradores en [12](pp:435) afirman, dándolo por hecho de todos conocido, que una gramática formal es, simplemente, un sistema deductivo formado por axiomas y reglas de inferencia, pero fallan en describirlo y en relacionarlo con otros sistemas deductivos conocidos, mientras que Deransart y colaboradores en [8] intentan el camino contrario; relacionar la lógica con las gramáticas. La formalización más completa, en este sentido y de la que tenemos conocimiento, se puede encontrar en unas cuantas líneas de Lambek [10], aunque nosotros seguiremos nuestro propio sistema.

Aunque la ecuación, *gramática=sistema deductivo*, es perfectamente válida, creemos que ha oscurecido un poco la naturaleza de las GIC. Proponemos otro enfoque, el de *gramática=relación de orden parcial*, que aunque equivalente al anterior (un sistema deductivo en el que se admite la regla del corte establece una relación de orden parcial) esperamos sea más iluminador ya que permite observar con más claridad que se suelen manejar dos sistemas deductivos superpuestos: la gramática en sí, y el mecanismo para el análisis sintáctico.

Dada una gramática independiente de contexto formada por la tupla: $\{T, N, S, P\}$, donde:

- T es el conjunto de tokens (símbolos terminales).

- N es el conjunto de símbolos no terminales.
- S es un elemento distinguido del conjunto N (el símbolo inicial).
- P es un conjunto de reglas de la forma: $A \rightarrow \alpha$. con $\alpha \in (T \cup N)^*$
- Podemos admitir un token especial, ξ, que no tiene representación externa.

Proponemos la siguiente lectura:

1. Formemos el conjunto de categorías sintácticas básicas, BASCAT $= T \cup N \cup \{\bot, \top\}$, y el conjunto BASCAT$^+$ mediante la clausura inductiva de BASCAT bajo la operación de concatenación $(\cdot)$ que podemos suponer asociativa, y dotemos a este conjunto de una relación de orden parcial tal y como está descrito en la sección anterior.

2. Las reglas de una GIC de la forma: $A \rightarrow x_1\ x_2\ \ldots\ x_n$ hay que entenderlas así: $A \geq ((((x_1 \cdot x_2) \cdot x_3) \cdot \ldots) \cdot x_n)$

   y una regla de la forma: $A \rightarrow \alpha \mid \beta$, como dos restricciones diferentes: $A \geq \alpha$, y $A \geq \beta$. Estableciendo una relación de orden entre ambas categorías (relación debida a las restricciones gramaticales). Es decir, estableciendo una teoría que se superpone al orden parcial de la lógica subyacente que poseíamos antes.

   Al orden definido por esta lectura de las reglas de una GIC le llamaremos *orden de las restricciones jerárquicas*, y refleja conocimientos concretos de cada lenguaje.

3. Ahora definimos el espacio de las categorías sintácticas $\mathcal{CAT}$ como el conjunto de los ideales$_\wedge$ del conjunto BASCAT$^+$ bajo la relación de orden definida por: el orden de la lógica subyacente, y por el orden de las restricciones jerárquicas. El conjunto $\mathcal{CAT}$ está ordenado bajo la relación de inclusión y tiene estructura de semiretículo$_\wedge$.

Aunque en una gramática independiente de contexto no se suelen admitir más limitaciones sobre las formas de las reglas gramaticales que el que la cabecera esté formada por un único símbolo no terminal, y el cuerpo por una cadena de símbolos terminales y no terminales; con estas únicas condiciones se pueden escribir gramáticas sin sentido lingüístico.

Cabe preguntarse: ¿Qué condiciones deben exigirse a una GIC para que ésta tenga una posible realización lingüística? (entendiendo por ello que describa algún lenguaje, bien sea natural o formal, que se pueda utilizar para comunicar alguna información). Nuestra conjetura es que:

- La teoría establecida por las restricciones jerárquicas defina una relación de orden parcial compatible con el orden de la lógica subyacente

de forma que la operación de concatenación siga siendo isótona y que se obtenga la estructura de semiretículo$_\wedge$.

- También deben cumplirse los requisitos llamados *regla de partición* y *regla de aplicación directa e inversa* explicados en la sección siguiente.

De ser esta conjetura cierta nos permitiría caracterizar, de una manera sencilla e intuitiva, a aquellas gramáticas que son lingüísticamente significativas; lo cual es mucho más difícil de conseguir utilizando exclusivamente la notación de las GIC. Es decir, el formalismo de las GIC resulta demasiado potente. Hay que notar que una buena teoría no debe admitir cualquier fenómeno si no que debe abarcar, de una manera simple, a aquellos fenómenos que son significativos y rechazar los que no lo son.

Esta teoría puede ser ampliada a las gramáticas sensibles al contexto pero, en este caso, habría que tener muy en cuenta a la propiedad de isotonía de la operación de concatenacion para que la nueva teoría sea compatible con la lógica subyacente.

5 Fundamentación algebraica de las GC$_\leq$: Modelo algebraico para las GC$_\leq$

Ya hemos visto un algoritmo por el que, a partir de una gramática independiente de contexto (que cumple ciertos requisitos), se obtiene un conjunto de categorías, BASCAT$^+$, formado por la clausura inductiva del conjunto de categorías sintácticas básicas y la operación de concatenación. A partir de él se construye el conjunto $\mathcal{CAT}$ de categorías gramaticales constituido por los ideales$_\wedge$ del conjunto anterior. Falta relacionar este conjunto $\mathcal{CAT}$ con el que se suele utilizar en el formalismo de las gramáticas categoriales. Estas últimas utilizan las operaciones de división por la izquierda y por la derecha. El nexo lo establece una versión de la teoría de conjuntos residuados (véase por ejemplo [11]).

Recordemos que en el conjunto BASCAT$^+$, el ideal$_\wedge$ principal generado por una cierta categoría, x, contiene a todas las categorías que son compatibles con x pero que están más especificadas. Cuando afirmamos que una cierta cadena de símbolos pertenecen a dicha categoría, x, en realidad estamos indicando que hasta el momento actual tenemos poca información pero que, conforme vayamos añadiendo información, iremos descendiendo por las categorías pertenecientes al ideal$_\wedge$ principal generado por dicha categoría x. Con esta idea hemos definido el conjunto $\mathcal{CAT}$ formado por los ideales$_\wedge$ de BASCAT$^+$.

Este mecanismo de las restricciones gramáticales que nos hacen descender por los elementos de un conjunto ordenado obteniendo la información más y más definida, recuerda, en cierta manera, a la antigua idea filosófica de que el significado lo vamos obteniendo por exclusión de lo que no es.

En nuestro caso el resultado sería todo un conjunto de categorías que son compatibles con la información proporcionada y de donde se han excluido todas aquellas que son incompatibles.

Es bien conocido que las GC tienen una posible interpretación algebraica. La pregunta inmediata es: ¿Podríamos utilizar el mismo modelo algebraico, obtenido en la sección anterior para las GIC, como fundamentación algebraica de las GC?.

La respuesta es que SI PERO a cambio de restringirnos a aquel subconjunto de las GIC que cumplan los requisitos descritos más adelante en esta misma sección.

Ahora, siguiendo a Moortgat y Oehrle [11] construiremos las GC a partir del modelo algebraico anterior. Adviértase que la construcción difiere de la realizada en [11] debido a la estructura en forma de dominio de Scott de nuestro conjunto BASCAT$^+$.

DEFINICIÓN 1 (producto) *Definiremos así el producto ($\otimes$) en el conjunto de categorías $\mathcal{CAT}$:*

Dados $X, Y \in \mathcal{CAT}$, $X \otimes Y = \triangle(\{x \cdot y \mid x \in X,\ y \in Y\})$

Donde hemos utilizado la notacion: $Ideal_\wedge$ *generado por*$(\alpha) = \triangle(\alpha)$

Notar la diferencia que existe con la definición que se suele hacer del producto en las GC, véase por ejemplo Lambek [9] o Moortgat y Oehrle [11], y que es como sigue:

Dados $X, Y \in \mathcal{CAT}$, $X \otimes Y = \{x \cdot y \mid x \in X,\ y \in Y\}$

La diferencia estriba en que, en la definición habitual, el conjunto de categorías BASCAT$^+$ no tiene ninguna estructura interna mientras que en nuestro caso mantiene una estructura de dominio de Scott.

LEMA 1 (isotonía del producto) *El producto es isótono respecto a la relación de orden establecida en $\mathcal{CAT}$ mediante la inclusión. Es decir, si $B \subseteq C$, entonces $X \otimes B \subseteq X \otimes C$, y análogamente $B \otimes X \subseteq C \otimes X$.*

La demostración de la isotonicidad es como sigue: Supongamos dos ideales$_\wedge$ $Y, Y' \in \mathcal{CAT}$ tales que $Y \subseteq Y'$; entonces, como $\mathcal{CAT}$ está ordenado por la relación de inclusión, $\forall y \in Y$, se cumple que $y \in Y'$; de aquí podemos deducir que $\{x \cdot y \mid x \in X,\ y \in Y'\} \subseteq \{x \cdot y' \mid x \in X,\ y' \in Y'\}$, lo cual implica que $\triangle\{x \cdot y \mid x \in X,\ y \in Y\} \subseteq \triangle\{x \cdot y' \mid x \in X,\ y' \in Y'\}$; es decir, que $X \otimes Y \subseteq X \otimes Y'$; y análogamente para el producto por la derecha.

LEMA 2 (asociatividad del producto) *En el supuesto de que la concatenación ($\cdot$) en el conjunto BASCAT$^+$ sea asociativa, entonces el producto ($\otimes$) en el conjunto $\mathcal{CAT}$ tambien es asociativo.*

Es decir: $(X \otimes Y) \otimes Z = X \otimes (Y \otimes Z)$.

La demostración es como sigue: por la propia definición del producto de categorías $\forall(xy)z \in (X \otimes Y) \otimes Z \Rightarrow \exists xy \in X \otimes Y$, $\exists z \in Z$ tales que $(xy)z \leq xy \cdot z$.

Tambien por la definición del producto deducimos que: $\forall xy \in X \otimes Y \Rightarrow \exists x \in X$, $\exists y \in Y$ tales que $xy \leq x \cdot y$.

Ahora tendremos también en cuenta que el producto es isótono y asociativo: $(xy)z \leq xy \cdot z \leq (x \cdot y) \cdot z = x \cdot (y \cdot z) \in X \otimes (Y \otimes Z)$. Y al ser $X \otimes (Y \otimes Z)$ un ideal de aquí se deduce que $\forall(xy)z \in (X \otimes Y) \otimes Z \Rightarrow (xy)z \in X \otimes (Y \otimes Z)$, es decir que: $(X \otimes Y) \otimes Z \subseteq X \otimes (Y \otimes Z)$.

La demostración de $(X \otimes Y) \otimes Z \supseteq X \otimes (Y \otimes Z)$ es similar.

DEFINICIÓN 2 (funciones coma) *Ahora consideremos las dos siguientes familias de funciones entre los conjuntos de ideales$_\wedge$: Dado un elemento cualquiera, X, del conjunto $\mathcal{CAT}$, definiremos dos funciones*

$\lambda_X : \mathcal{CAT} \mapsto \mathcal{CAT}$
$\rho_X : \mathcal{CAT} \mapsto \mathcal{CAT}$
$\lambda_X(Y) = X \otimes Y$
$\rho_X(Y) = Y \otimes X$

LEMA 3 (isotonía de las funciones coma λ_X y ρ_X) *Estas funciones transforman ideales$_\wedge$ en ideales$_\wedge$ (por construcción), y además son isótonas respecto a la relación de orden definida en $\mathcal{CAT}$ por la inclusión. Es decir, si $B \subseteq C$, entonces $\lambda_X(B) \subseteq \lambda_X(C)$. Análogamente para ρ_X.*

La razón está en la isotonicidad del producto. Nos interesa ahora estudiar si estas funciones tienen algún tipo de inversa.

DEFINICIÓN 3 (funciones resíduo) *Dado un elemento cualquiera, X, de $\mathcal{CAT}$, consideremos las dos funciones:*

$\lambda_X^{\leftarrow} : \mathcal{CAT} \mapsto \mathcal{CAT}$
$\rho_X^{\leftarrow} : \mathcal{CAT} \mapsto \mathcal{CAT}$
Definidas así:
$\lambda_X^{\leftarrow}(Y) = \{z \in \text{BASCAT}^+ \mid \exists x \in X \ \ tal\ que\ x \cdot z \in Y\}$
$\rho_X^{\leftarrow}(Y) = \{z \in \text{BASCAT}^+ \mid \exists x \in X \ \ tal\ que\ z \cdot x \in Y\}$

LEMA 4 *Las imágenes, bajo estas funciones, de un ideal$_\wedge$ no pueden ser vacías.*

Por lo menos contienen al elemento $\perp$, ya que este pertenece a todos los ideales$_\wedge$; y para todo $x \in \text{BASCAT}^+$, se cumple que $x \cdot \perp = \perp \cdot x = \perp$.

LEMA 5 *Debemos también comprobar que las imágenes de un ideal$_\wedge$ son, efectivamente, un ideal$_\wedge$.*

Es decir, tomando un cierto $z \in \lambda_X^{\leftarrow}(Y)$, eso implica que $\exists x \in X$ tal que $x \cdot z \in Y$; entonces $\forall z'$ tal que $z' \leq z$, debemos demostrar que $z' \in \lambda_X^{\leftarrow}(Y)$.

Demostración: Por la isotonicidad del producto se debe cumplir que $x \cdot z' \leq x \cdot z$; y como $x \cdot z \in Y$, que es un ideal$_\wedge$, entonces también se debe cumplir que $x \cdot z' \in Y$; lo cual implica (por la definición) que $z' \in \lambda_X^{\leftarrow}(Y)$.

LEMA 6 (isotonicidad de las funciones $\lambda_X^{\leftarrow}$, y $\rho_X^{\leftarrow}$) *Las funciones $\lambda_X^{\leftarrow}$, y $\rho_X^{\leftarrow}$, conservan el orden. Es decir, si $B \subseteq C$, entonces $\lambda_X^{\leftarrow}(B) \subseteq \lambda_X^{\leftarrow}(C)$*

Demostración: tenemos que por definición $\lambda_X^{\leftarrow}(B) = \{z \mid \exists x \in X,\ x \cdot z \in B\}$. Entonces, si $B \subseteq C$, debe ocurrir que $\forall z \in \lambda_X^{\leftarrow}(B)$, $\exists x \in X$ tal que $x \cdot z \in B$, lo cual implica que $x \cdot z \in C$, lo que a su vez implica que $z \in \lambda_X^{\leftarrow}(C)$; es decir, que $\lambda_X^{\leftarrow}(B) \subseteq \lambda_X^{\leftarrow}(C)$, como queríamos demostrar.

DEFINICIÓN 4 (división de categorías) *Dados dos ideales$_\wedge$ $X, Y \in \mathcal{CAT}$; llamaremos $X\backslash Y$ al ideal$_\wedge$ $\lambda_X^{\leftarrow}(Y)$.*

Análogamente, denominaremos Y/X al ideal$_\wedge$ $\rho_X^{\leftarrow}(Y)$.

DEFINICIÓN 5 (funcion unidad) *Definimos una función unidad $1_{\mathcal{CAT}} : \mathcal{CAT} \mapsto \mathcal{CAT}$ que transforma a cada categoría $X \in \mathcal{CAT}$ en sí misma $1_{\mathcal{CAT}}(X) = X$. Esta función tiene las propiedades:*

$\lambda_X \circ 1_{\mathcal{CAT}} = \lambda_X = 1_{\mathcal{CAT}} \circ \lambda_X$,
$\rho_X \circ 1_{\mathcal{CAT}} = \rho_X = 1_{\mathcal{CAT}} \circ \rho_X$,
$\lambda_X^{\leftarrow} \circ 1_{\mathcal{CAT}} = \lambda_X^{\leftarrow} = 1_{\mathcal{CAT}} \circ \lambda_X^{\leftarrow}$,
$\rho_X^{\leftarrow} \circ 1_{\mathcal{CAT}} = \rho_X^{\leftarrow} = 1_{\mathcal{CAT}} \circ \rho_X^{\leftarrow}$,

En el caso de que admitamos la categoría $\xi \in BASCAT$ como aquella que no tiene representación externa (la cadena vacía) y que es el elemento neutro para la concatenación, y suponiendo que las restricciones jerárquicas son tales que no permiten a ξ en el lado izquierdo de una regla gramatical; entonces el ideal$_\wedge$ generado por ξ forma una categoría sintáctica que llamaremos $\epsilon \in \mathcal{CAT}$ tal que $1_{\mathcal{CAT}} = \lambda_\epsilon = \rho_\epsilon$.

REQUISITO 1 (Regla de partición, splitting) *Pediremos que se cumpla que:* $(\lambda_X^{\leftarrow} \circ \lambda_X) = 1_{\mathcal{CAT}}$, $(\rho_X^{\leftarrow} \circ \rho_X) = 1_{\mathcal{CAT}}$

O lo que es equivalente: $X\backslash(X \otimes Y) = Y$; $(Y \otimes X)/X = Y$

Exigiremos que la forma de las restricciones gramaticales sea tal que la estructura resultante cumpla esta regla. [1]

Así ocurre en el caso de la mayoría de las gramáticas independiente de contexto que se utilizan en la práctica pero, en caso de tratarse de una gramática sensible al contexto, este lema no se cumpliría en general sino que habría que imponerlo restringiendo la forma de las reglas jerárquicas. También podría servir de criterio para clasificar la complejidad de las gramáticas.

[1] Notar que hemos utilizado la siguiente notación para la composición de funciones: $(f \circ g)(x) = f(g(x))$.

REQUISITO 2 (Reglas de aplicación directa e inversa) *pediremos que:* $(\lambda_X \circ \lambda_X^{\leftarrow}) = 1_{\mathcal{CAT}}$, $(\rho_X \circ \rho_X^{\leftarrow}) = 1_{\mathcal{CAT}}$

O equivalentemente: $X \otimes (X\backslash Y) = Y$; $(Y/X) \otimes X = Y$

También pediremos que la forma de las restricciones gramaticales sea tal que se cumpla esta regla.

Existen muchas propiedades generales de la división que se utilizan en las gramáticas categoriales y que se pueden deducir del modelo anterior, constituyendo una fundamentación algebraica de la teoría de gramáticas categoriales.

LEMA 7 (Reglas de introducción de la división) *Si* $X \otimes Z \subseteq Y$ *entonces* $Z \subseteq X\backslash Y$

O equivalentemente: Si $Z \otimes X \subseteq Y$ *entonces* $Z \subseteq Y/X$

La demostración es inmediata utilizando el requisito 1 anterior (regla de partición): $\lambda_X^{\leftarrow} \circ \lambda_X = 1_{\mathcal{CAT}}$; así como el lema 6 (propiedad de isotonicidad de la función $\lambda_X^{\leftarrow}$).

En efecto: si $X \otimes Z = \lambda_X(Z) \subseteq Y$; aplicando ahora la función $\lambda_X^{\leftarrow}$ (que es isótona) se debe cumplir que: $\lambda_X^{\leftarrow}(\lambda_X(Z)) = (\lambda_X^{\leftarrow} \circ \lambda_X)(Z) = Z \subseteq \lambda_X^{\leftarrow}(Y) = X\backslash Y$.

LEMA 8 (Regla de elevación de tipos) $X = Z/(X\backslash Z)$; $X = (Z/X)\backslash Z$

Para la demostración hay que notar que por el requisito 1 (regla de partición): $\rho_K^{\leftarrow} \circ \rho_K = 1_{\mathcal{CAT}} \Leftrightarrow (Y \otimes K)/K = Y$, y que por el requisito 2 (regla de aplicación inversa): $\rho_{X\backslash Z}(X) = X \otimes (X\backslash Z) = Z$. Entonces: $X = 1_{\mathcal{CAT}}(X) = (\rho_{X\backslash Z}^{\leftarrow} \circ \rho_{X\backslash Z})(X) = \rho_{X\backslash Z}^{\leftarrow}(\rho_{X\backslash Z}(X)) = \rho_{X\backslash Z}^{\leftarrow}(Z) = Z/(X\backslash Z)$

Análogamente para la otra fórmula.

Dejaremos para otros artículos la demostración del resto de las reglas habituales en las gramáticas categoriales, así como el estudio de las posibles variaciones y parametrizaciones de la nueva teoría y, sobre todo, el papel del análisis sintáctico en esta nueva extensión.

6 Conclusiones

Hemos descrito una extensión a las gramáticas categoriales $GC_{\leq}$, que nos permite pasar facilmente de una descripción de un lenguaje basada en GIC a otra basada en GC y que, lo más importante, nos mantiene la semántica (el modelo algebraico común). Esta extensión surgió al estudiar la relación entre ambas teorías.

Durante este estudio hemos encontrado dos razones que hacen que que el formalismo de las gramáticas categoriales sea más débil que el de las gramáticas independientes de contexto:

- Existe un fenómeno de *jerarquía* que está presente en las GIC pero que falta en las GC. Sin embargo se puede obtener el mismo efecto en las GC por el mecanismo de definir una relación de orden parcial en el espacio de categorías sintácticas, tomar los ideales$_\wedge$ como los elementos de un nuevo espacio, y las reglas de una GIC como relaciones de orden entre categorías. Así se refleja la naturaleza *incompleta*, *basada en restricciones* y *basada en herencia* de la información lingüística.

- Las GIC son más potentes que las GC debido a que las primeras pueden ser utilizadas para definir lenguajes sin ningún sentido lingüístico. Es decir, las GIC no discriminan entre gramáticas significativas y gramáticas no–significativas. Pero esto no es una característica deseable de un formalismo. Lanzamos la conjetura (que debe ser confirmada o rechazada por estudios posteriores) de que las GC$_\leq$ (GC aumentadas con una relación de orden parcial) describen, a un subconjunto de las GIC y GSC que son *lingüísticamente significativas*.

Tambien hemos desarrollado un modelo matemático común a cierto subconjunto de las GIC y a las GC$_\leq$ comprobando, de esta forma, la equivalencia en poder generativo fuerte de ambos formalismos. Este modelo puede ser interpretado como una fundamentación algebraica de las GC$_\leq$, con una clara inspiración intuitiva. Pero algunas de las propiedades deben ser impuestas (y no deducidas del modelo). Esto está de acuerdo con la idea de que las GIC son un formalismo demasiado potente y que debe ser limitado. Solo deberíamos tener en cuenta aquellas gramáticas que cumplan ciertos requisitos.

La relación de orden está definida mediante dos teorías diferentes: Una teoría de la lógica subyacente a la operación de concatenación de estructuras sintácticas; y otra teoría jerárquica que refleja conocimientos sintácticos específicos de la lengua concreta que estemos considerando.

Finalmente apuntamos la posibilidad de construir un modelo único a todas las gramáticas sintagmáticas significativas (cualquiera que sea el formalismo gramatical en que estén escritas). Este modelo vendría parametrizado por las propiedades matemáticas de la estructura algebraica del espacio de categorías, dando un paso más en la dirección de desarrollar una teoría lógico–algebraica avanzada de las gramáticas formales.

Referencias

[1] Hassan Aït-Kaci, Richard Meyer, Peter Van Roy, y Bruno Dumant. *"Wild-Life 1.0 User Manual (Draft) Proteus Project"*, Digital Equipment Corporation, 1993.

[2] John Backus. *"Programming in America in the 1950s-Some Personal Impressions"*. En: N. Metropolis, J. Howlett, y Gian-Carlo Rota, eds.

"The History of Computing in the Twentieth Century. A Collection of Essays with Introductory Essay and Indexes", Academic Press, 1980.

[3] Y. Bar–Hillel, C. Gaifman, y E. Shamir. *"On Categorial and Phrase Structure Grammars"* Bulletin Research Council of Israel, F9, 1960.

[4] Wojciech Buszkowski. *"Generative Power of Categorial Grammars"*. En: Emmon Bach y Dreire Wheeler *"Categorial Grammars and Natural Languaje Structures"*. Reidel Publishing Company, 1988.

[5] Noam Chomsky. *"Formal Properties of Grammars"*. En: R. Duncan-Luce y Col. (Eds). *"Handbook of Mathematical Psychology, Vol.2"*. Willey, 1963.

[6] Noam Chomsky. *"Aspects of the Theory of Syntax"*. The M.I.T. Press, 1965.

[7] Walter Daelemans, Koenraad De Smedt, y Gerald Gazdar. *"Inheritance in Natural Language Processing"*, Computational Linguistics, Vol. 18, Number 2, 1992.

[8] Pierre Deransart, y Jan Maluszynski. *"What kind of Grammars are Logic Programs?"*. En: Patrick Saint-Dizier, y Stan Spakowicz. *"Logic and Logic Grammas for Language Processing"*, Ellis Horwood, 1990.

[9] Joachim Lambek. *"Categorial and Categorical Grammars"*. En: R.T. Oehrle, E. Bach, y Deirdre Wheeler Eds. *"Categorial Grammars and Natural Language Structures"*, D. Reidel Publishing Company, 1988.

[10] Joachim Lambek. *"Logic Without Structural Rules (Another Look at Cut Elimination)"*. En Peter Schoreder-Heister y Kosta Dosen Eds. *"Substructural Logics"* Clarendon Press–Oxford, 1993.

[11] Michael Moortgat, y Dick Oehrle. *"Categorial Grammar: Logical Parameters and Linguistic Variation"*, Readings from the Fith European Summer School in Logic Language and Information. Lisboa, 1993.

[12] Barbara H. Partee, Alice Ter Meulen, y Robert E. Wall. *"Mathematical Methods in Linguistics"*, Kluwer Academic Publishers, 1993.

[13] Mati Pentus. *"Lambek Grammars are Context Free"*. En Proceedings of LICS, 1992.

[14] Fernando C.N. Pereira, y David H.D. Warren. *"Parsing As Deduction"*, Proceedings of the Conference of the 21st Annual Meeting of the Association for Computational Linguistics. M.I.T. 1983.

[15] Carl Pollard, e Ivan A. Sag. *"Information–Based Syntax and Semantics. Vol 1: Fundamentals"*, CSLI Lecture Notes N. 13. 1987.

[16] Dana S. Scott. *"Domains for Denotational Semantics"*, en M. Nielsen, y E.M. Schmidt, Eds. *"Automata, Languages and Programming, Ninth Colloquium"*, Lecture Notes in Computer Science, Vol. 140. Springer Verlag, 1982.

[17] Stuart M. Shieber. *"Constraint-Based Grammar Formalisms: Parsing and Type Inference for Natural and Computer Languages"*, M.I.T. Press, 1992.

[18] J. Van Benthem. *"Language in Action: Categories, Lambdas and Dynamic Logic"*. North–Holland, 1991.

[19] Y. Wilks. *"Does Anyone Really Still Believe This Kind of Things?"*. En Karen Sparck Jones, Y. Wilks Eds. *"Automatic Natural Language Parsing"*. Ellis Horwood Limited, 1985.