INFORMACIÓN

NOTA SOBRE LOS CONCEPTOS DE SISTEMA Y LENGUA FORMALES

1. Abordar el estudio de una parcela de la realidad científicamente tiene un sentido bastante preciso desde el siglo XVII. Son rasgos característicos la o b j e t i v i d a d que se apoya en la observación y en la medida (en el que el instrumento de medida es de capital importancia) y la l ó g i c a, entendida especialmente como deductiva (en la que el algoritmo juega un papel importante). Es claro que algunas áreas de conocimiento son más aptas para ser abarcadas por un estudio científico, por ser más sencillas, más regulares, más identificables las magnitudes que las componen y sus relaciones entre ellas. Así el movimiento, el calor, la luz, la electricidad, etc.... han sido estudiados científicamente y construidas las ciencias correspondientes. Puede parecer ahora sencillo y especialmente adecuado el estudio de esas ciencias, pero vagamente planteados los problemas de las mismas no dejan de ser de envergadura, o también si su planteamiento se hiciese en términos absolutos pretendiendo abarcar el s e r de la luz, calor o movimiento, y no se hubiera restringido su estudio a sólo unos aspectos relativos y en función de una operatividad y finalidad específicas.

Frente a los aspectos «naturales» de la realidad se suelen contraponer los aspectos «humanos». Frente a las Ciencias Naturales las Ciencias Humanas, que se pretende que no sólo difieren de objeto, sino también de método, cuando la historia de la Ciencia nos muestra que el método no depende tanto del objeto cuanto de la situación histórica y social de dicho objeto.

2. Un área de conocimiento especialmente delicada es el lenguaje, ya que está en el entronque mismo de todo conocimiento humano. En este área es más difícil la objetividad, pues siempre el sujeto está presente, está entrelazado con el lenguaje, no son fáciles de definir magnitudes y, por tanto, construir instrumentos de medida. Tampoco es sencillo construir una lógica deductiva

del lenguaje humano, tal vez sea imposible. Por todo esto a veces se ha considerado el lenguaje como apto sólo de ser estudiado dentro de ese vago concepto de Ciencias Humanas. Sin embargo, por impulso fundamentalmente de la lógica, desde hace más de un siglo se persigue la «formalización del lenguaje», aunque por este camino lo más que se ha conseguido es la construcción de «lenguajes formales».

3. Desde un punto de vista formal vamos a considerar un lenguaje simplemente como un conjunto finito o infinito de cadenas construidas con elementos tomados de un alfabeto finito Σ. Es decir, partimos de un conjunto finito de símbolos

$$\Sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_n\}$$

denominado *alfabeto* (también a veces *vocabulario)* y con él construimos cadenas de la siguiente forma:

1. Designamos por Λ a la cadena que no tiene ningún elemento y que llamamos «cadena vacía».
2. Si x es una «cadena», entonces $\forall \sigma \in \Sigma$ se tiene que $x\sigma$ es una cadena.
3. Sólo son cadenas las obtenidas mediante 1 y 2.

Es decir, en general una cadena x la podemos escribir así:

$$x = \sigma_{i_1}, \sigma_{i_2}, \ldots, \sigma_{i_k}$$

donde $\sigma_{i_j} \in \Sigma$ para $j = 1, 2, \ldots, k$. (A veces a las cadenas las llamaremos indistintamente «palabras»).

Llamaremos «lenguaje universo» generado por el alfabeto Σ al conjunto de todas las cadenas que se pueden formar con las letras de dicho alfabeto. Al lenguaje universo lo denotamos por Σ^*, y evidentemente tiene un número infinito numerable de cadenas, es decir, coordinable con la serie de los números naturales.

En Σ^* podemos definir una operación binaria (es decir, una operación que a cada dos palabras de Σ^* se le asocia como resultado una palabra de Σ^*) llamada «concatenación», mediante la cual a dos cadenas arbitrarias x, y $\in \Sigma^*$ se le asocia otra cadena $z \in \Sigma^*$, obtenida por yuxtaposición de x e y. Es decir, que si

$$x = x_1\, x_2 \ldots x_p \qquad x_i \in \Sigma \quad i = 1, 2, \ldots p$$
$$y = y_1\, y_2 \ldots y_p \qquad x_j \in \Sigma \quad j = 1, 2, \ldots q$$

entonces la cadena $z = xy$ sería

$$z = x_1 x_2 \ldots x_p y_1 y_2 \ldots y_q$$

Como evidentemente la concatenación es asociativa, es decir cumple que

$$(xy)z = x(yz)$$

y la cadena vacía Λ es neutro de la concatenación, es decir, cumple para todo $x \in \Sigma^*$ que

$$x\Lambda = \Lambda\, x = x$$

el sistema

$$\{\Sigma^*,\ \text{concatenación}\}$$

posee la estructura algebraica de semigrupo con unidad o monoide. Frecuentemente se designa a Σ^* como el monoide Σ^*.

Visto esto, llamaremos «lenguaje» L a todo subconjunto de Σ^*, es decir si

$$L \subset \Sigma^*$$

decimos que L es un lenguaje en Σ. Diremos que L es finito o infinito si está compuesto por un número finito o infinito de cadenas de Σ^*. Al lenguaje que sólo contiene la cadena vacía Λ se le denota por L_Λ y al lenguaje que no contiene ninguna cadena se le llama «lenguaje vacío» y se le denota por L_ϕ. Obsérvese que

$$L_\Lambda \neq L_\phi$$

Dados los lenguajes L_1 y L_2 en el alfabeto Σ, se pueden definir diferentes operaciones algebraicas, entre las que, a modo de ejemplo, podemos citar *:

UNIÓN: $L_1 \cup L_2 = \{x | x \ L_1 \vee x \ L_2\}$

INTERSECCIÓN: $L_1 \cap L_2 = \{x | x \ L_1 \wedge x \ L_2\}$

COMPLEMENTO: $\sim L_1 = \{x | x \ \Sigma^* \wedge x \ L_1\}$

PRODUCTO: $L_1 \,.\, L_2 = \{z | z = xy \wedge x \, L_1 \wedge y \in L_2\}$

CLAUSURA: $L_1^* = \bigcup_{1=o}^{\infty} L_1^i$ donde $L_1^o = L_\Lambda$ y $L_1^i = L_1^{i-1} \,.\, L_1$

4. Para introducir la idea de gramática formal, introduciremos antes la noción más general de sistema formal. Definiremos un sistema formal como un procedimiento para construir deductivamente los teoremas de una teoría, considerados estos teoremas en su forma lingüística, o mejor, en su forma sintác-

---

* Las notaciones empleadas son las usuales en el álgebra de conjuntos y en el cálculo de proposiciones; así la barra vertical | se lee «tal que», el símbolo $\in$ indica la pertenencia a un conjunto, es decir, la expresión $x \in L_1$ se lee «*x pertenece a* L_1» y la expresión $x \notin L_1$ se lee «*x no pertenece a* L_1»; los símbolos $\vee$, $\wedge$ indican el «o inclusivo» y la «conjunción», respectivamente. Los símbolos $\cup$, $\cap$, $\sim$, ., * son con los que indicamos las operaciones definidas;

$$\bigcup_{i=o}^{\infty} L_1^i \text{ denota } L_\Lambda \cup L \cup L.L \cup L.L.L \cup \ldots$$

tica, y sin tener en cuenta sus significados o las interpretaciones que de ellos se hagan; por tanto, se trata de una manera de caracterizar ciertos tipos de lenguajes, especialmente orientados hacia la lógica o la fundamentación de la matemática, pero que pueden también emplearse para el estudio de algunos aspectos de la sintaxis de los lenguajes naturales.

Sintéticamente diremos que un sistema formal S está constituido por cuatro elementos:

$$S = (\Sigma , F , A , P)$$

donde Σ es un alfabeto; F un subconjunto recursivo (*) de Σ^* llamado conjunto de fórmulas, ($F \subset \Sigma^*$); A es un subconjunto recursivo de F, cuyos elementos reciben el nombre de *axiomas*, ($A \subset F$); y P es un conjunto finito de predicados recursivos (de orden mayor que 1), cuyas variables toman valores en F. Los elementos de P se denominan «reglas de inferencia».

En una regla de inferencia $p \in P$, tal que $p(x_1, x_2, ..., x_n, y)$, se dice que *y* se obtiene a partir de $x_1, x_2, ..., x_n$ mediante p.

Para dar la noción de teorema formal, veamos antes qué entendemos por demostración. Dado un sistema formal S, decimos que *d* es una demostración, si es una secuencia finita de fórmulas

$$x_1, x_2, ..., x_n$$

tales que para todo $i \leqslant n$ se verifica una de las siguientes condiciones:

1) $x_i \in A$, es decir, es un axioma.
2) existe una subsecuencia de fórmulas tomadas de la secuencia que constituye la demostración

$$x_{j_1}, x_{j_2}, ..., x_{j_k}$$

con $k < i$, tal que

$$p(x_{j_1}, x_{j_2}, ..., x_{j_k}, x_i) \in P$$

Informalmente, podríamos decir que una demostración es una sucesión de fórmulas que comienza con un axioma y en la que cualquier elemento o es un axioma o se puede deducir (inferir) a partir de otros elementos que ya estuvieren incluidos en la demostración.

---

* La importante noción de recursividad está vinculada con la posibilidad de acceder mediante un procedimiento finito a cualquier elemento de un conjunto infinito. Funciones recursivas son las que se pueden calcular para cualquier argumento aplicando un número finito de veces algunas de las operaciones llamadas «de base», un conjunto es recursivo cuando su función característica es recursiva, y un predicado es recursivo cuando su extensión (es decir, el conjunto de elementos que lo cumplen) es recursiva.

A partir de la definición de demostración, llamaremos *teorema* a toda fórmula que ocupa último lugar en una demostración. En otras palabras, diremos que la fórmula x_n es un teorema, si y sólo si existe una demostración

$$d = x_1, x_2, \ldots, x_n$$

El conjunto de todos los teoremas de un sistema formal constituyen un lenguaje en el sentido empleado más arriba.

5. La gran utilidad práctica de los sistemas formales se vio ampliamente en el presente siglo, fundamentalmente en lo que respecta a la lógica, a la matemática y a las ciencias en que éstas se aplican directamente. La axiomatización de las teorías, la búsqueda de estructuras abstractas y generales ha predominado en el desarrollo actual de las ciencias. Y aunque los teoremas limitadores de Gödel y Church han puesto ciertas fronteras al entusiasmo, no deja de constituir el soporte sólido en el que apoyarse cuando se trata de construir procedimientos efectivos, es decir, procedimientos en los que con la aplicación de un número finito de reglas tomadas de un conjunto reducido de ellas se obtengan los resultados deseados; en definitiva, de procedimientos que podrían llamarse automáticos para la solución de los problemas llamados computables. Todas las modernas teorías de máquinas lógicas (Turing, Post, Kleene, etc....), de las que son tosca realización física los actuales ordenadores electrónicos, tendrían ese fundamento.

6. En particular algunos aspectos de la lingüística han sido formalizados en este sentido. Daremos a continuación la idea general de gramática formal, también llamada gramática generativa en cuanto que su finalidad consiste en la generación de todas las cadenas (palabras, frases, textos...) de un lenguaje determinado. Se trata de la caracterización de los lenguajes para que puedan enumerar sus cadenas mediante procedimientos finitos y efectivos; en otras palabras, de disponer de un número finito de reglas mediante las cuales podamos generar todas las cadenas de un lenguaje.

Formalmente diremos que una gramática es un sistema formal

$$G = (V, V^*, S, P)$$

donde V es un alfabeto que se descompone en dos subalfabetos disjuntos V_N y V_T, es decir

$$V = V_N \cup V_T \quad ,, \quad V_N \cap V_T = \emptyset$$

los alfabetos V_N y V_T reciben los nombres de vocabularios terminal y no terminal respectivamente. El segundo elemento V* será en nuestro caso el conjunto de fórmulas formado por todas las cadenas de cualquier longitud con letras de V. El tercer elemento S será el axioma único de nuestro sistema formal denominado generalmente «símbolo inicial» o de partida. Por último, el elemento P es el conjunto de los predicados que componen las reglas de infe-

rencia que en el caso particular de las gramáticas suelen denominarse «reglas de producción» o de «reescritura». Los predicados de P son de segundo orden de la forma

$$p(\alpha_i, \alpha_j) \qquad \alpha_i, \alpha_j \in V^*$$

y que podría leerse como α_i puede ser reescrito por α_j. El lenguaje generado por G sería el conjunto de sus *teoremas* terminales (entendidos éstos en el sentido formal que dimos más arriba).

La manera más habitual de describir las gramáticas formales suele ser la siguiente, que es fácilmente identificable con la anterior enunciación:

$$G = (V_N, V_T, S, P)$$

donde los predicados de P se escriben de la siguiente forma:

$$\alpha_i \rightarrow \alpha_j ;$$

donde α_i es una cadena no vacía compuesta por letras del alfabeto V, y α_j es una cadena de letras del mismo alfabeto.

Para generar una cadena mediante la gramática G, hemos de proceder del siguiente modo:

1.º Tomar una regla de P de la forma

$$S \rightarrow \alpha_1$$

2.º Pasamos de una cadena γ_i a otra γ_{i+1} ($\gamma_i, \gamma_{i+1} \in V^*$) mediante una regla de P, si

$$\left.\begin{array}{l} \gamma_i = \delta\ \alpha_i \delta' \\ \gamma_{i+1} = \delta\ \alpha_j \delta' \end{array}\right\} \delta\ ,\ \delta' \in V^*$$

y la regla $\alpha_i \rightarrow \alpha_j \in P$. En este caso escribiremos

$$\gamma_i \underset{G}{\Rightarrow} \gamma_{i+1}$$

3.º Si $\gamma_1 = \alpha_1$, y γ_m (obtenida aplicando *m* veces el procedimiento anterior) pertenece a V_T^*, diremos que γ_m es una cadena generada por G.

Al paso de S a γ_m se le denomina *derivación*, y se puede expresar

$$S \underset{G}{\Rightarrow} \gamma_1 \underset{G}{\Rightarrow} \gamma_2 \underset{G}{\Rightarrow} \cdots \underset{G}{\Rightarrow} \gamma_m$$

o simplemente

$$S \overset{*}{\underset{G}{\Rightarrow}} \gamma_m$$

que podría leerse: γ_m se obtiene por derivación de S aplicando reglas de G.

Podemos sintetizar lo anterior diciendo que el lenguaje generado por la gramática G es:

$$L(G) = \{x | x \in V_T^* \wedge S \overset{*}{\underset{G}{\Rightarrow}} x\}$$

Diremos que dos gramáticas son «equivalentes» si generan el mismo lenguaje, es decir, la gramática G_1 es equivalente a la G_2 si se verifica que

$$L(G_1) = L(G_2)$$

Ejemplo:

En la gramática

$$G = (\{S, A, B, C\}, \{0,1\}, S, P)$$

con

$$P = \{S \to AB,\ A \to CA,\ B \to BC,\ A \to 0,\ B \to 0,\ C \to 1\}$$

dos derivaciones serían

$$S \Rightarrow AB \Rightarrow CAB \Rightarrow CCAB \Rightarrow 1CAB \Rightarrow 11AB \Rightarrow 110B \Rightarrow 1100$$
$$S \Rightarrow AB \Rightarrow CAB \Rightarrow CABC \Rightarrow CCABC \Rightarrow CCABCC \Rightarrow 1CABCC \Rightarrow$$
$$11ABCC \Rightarrow 110BCC \Rightarrow 1100CC \Rightarrow 11001C \Rightarrow 110011$$

o sea que las palabras 1100 y 110011 pertenecen al lenguaje generado por G, es decir

$$1100 \in L(G)$$
$$110011 \in L(G)$$

Las palabras de L(G) tendrán la forma general

$$1^* 001^*$$

7. Las gramáticas generativas han sido jerarquizadas por Chomsky, según la forma particular de las reglas de producción contenidas en ella. Así, si tenemos una gramática

$$G = (V_N, V_T, P, S)$$

la llamaremos de tipo 0, tipo 1, tipo 2 y tipo 3, según las siguientes definiciones:

Gramáticas de tipo 0: Decimos que G es de tipo 0 si toda regla $\alpha_i \to \alpha_j \in P$ verifica que α_i es una cadena no vacía de símbolos tomados del alfabeto $V(V = V_N \cup V_T)$ y α_j es una cadena de símbolos de V. Es decir, la definición más general de gramática formal coincide con las llamadas de tipo 0.

Gramáticas de tipo 1: La gramática G es de tipo 1 si las reglas de producción $\alpha_i \to \alpha_j \in P$ cumplen la condición de que la longitud de la cadena α_i es siempre menor o igual que la longitud de la cadena α_j; siendo α_i y α_j, como

antes, cadenas formadas con símbolos de V. Otra forma de caracterizar este tipo de gramáticas es exigiendo que las reglas sean de la forma

$$\gamma_1 A \gamma_2 \rightarrow \gamma_1 \beta \gamma_2$$

donde γ_1, γ_2, $\beta \in V^*$, $\beta \neq \Lambda$, $A \in V_N$, que podría interpretarse como que la presente regla cambia el símbolo A por la cadena β siempre que A aparezca en el contexto $\gamma_1 - \gamma_2$. Por eso a estos gramáticas se las denomina también de «contexto sensitivo» *(context sensitive)*.

Gramáticas de tipo 2: La gramática G es de tipo 2 cuando las reglas de producción $\alpha_i \rightarrow \alpha_j \in P$ cumplen la condición de que α_i sea un único símbolo del alfabeto V_N, y α_j sea una cadena no vacía del alfabeto V ($V = V_N \cup V_T$); lo que puede interpretarse como que la regla $\alpha_i \rightarrow \alpha_j$ ordena la sustitución del símbolo no terminal α_i por la cadena α_j, independientemente del contexto en que aparezca α_i; por eso a estas gramáticas se las denomina también de «contexto libre» *(context free)*.

Gramáticas de tipo 3: La gramática G es de tipo 3 cuando las reglas de producción $\alpha_i \rightarrow \alpha_j \in P$ cumplen que α_i es un símbolo del alfabeto V_N y la cadena α_j o bien es un solo símbolo del alfabeto V_T o una cadena formada por un símbolo de V_T seguido de un símbolo de V_N. Estas gramáticas también reciben el nombre de «regulares».

Esta jerarquía de gramáticas induce de forma natural una jerarquía entre lenguajes, simplemente si llamamos a un lenguaje de tipo n ($n = 0, 1, 2,3$) cuando es generado por una gramática de tipo n, es decir, decimos que L es de tipo *n* si existe una gramática G de tipo n tal que

$$L = L(G)$$

Como las reglas de producción de cada tipo de gramática cumplen propiedades restrictivas sobre las reglas de tipo anterior, se tiene que, si llamamos C_0, C_1, C_2, C_3 a las clases de los lenguajes de tipo 0, de tipo 1, de tipo 2 y de tipo 3, respectivamente, podemos escribir:

$$C_3 \subset C_2 \subset C_1 \subset C_0$$

donde todas las inclusiones son propias. Obsérvese que todo lenguaje perteneciente a C_i ($i = 0, 1, 2, 3$) puede generarse por una gramática de tipo j tal que $j \leqslant i$.

8. Hemos visto las definiciones generales de una caracterización de los lenguajes, entendidos estos como colecciones bien definidas de cadenas formadas con un alfabeto particular. Esta caracterización evidentemente no es única y en ella se atiende principalmente a la capacidad generadora de las gramáticas, en un sentido muy informal a la capacidad de un «hablante» para producir frases:

Entre otras caracterizaciones se destaca la que atiende principalmente a discernir si una cadena dada pertenece o no a un lenguaje determinado, es decir, en términos informales, se sintetizaría la capacidad de un «oyente». El meca-

nismo formal asociado se denomina en sentido genérico «autómata» (a veces se utilizan otros nombres como «aceptadores», «sistemas de transición», «máquinas», etc....). La jerarquía que presentamos entre los lenguajes tiene su correlativa entre los autómatas, y a este respecto sólo diremos que los lenguajes de tipo 0 son reconocidos por las Máquinas de Turing, las de tipo 1 por los autómatas lineales acotados, las de tipo 2 por los autómatas a pilas y las de tipo 3 por los autómatas finitos.

9. La literatura escrita sobre estos temas es actualmente superabundante. Citaremos sólo a título indicativo algunos tratados que unos casos podrían considerarse como clásicos y en otros como obras de iniciación:

Bar-Hillel, Y., *Language and Information*, Addison Wesley, 1964.
Chomsky, N., «Formal Properties of Grammars», *Handbook of Mathematical Psichology*, John Wiley & Sons, 1963, cap. 12, págs. 323-418.
Chomsky, N. + Miller, G. A., «Introduction to the Formal Analysis of Natural Languages», *Handbook of Mathematical Psichology*, John Wiley & Sons, 1963, cap. 11, págs. 269-321.
Davis, M., *Computability and Unsolvability*, McGraw-Hill, 1958.
Ginsburg, S., *The Mathematical Theory of Context-Free Languages*, McGraw-Hill, 1966.
Gross, M. + Lentin, A., *Notions sur les grammaires formelles*, Gauthier-Villars, 1970.
Hermes, H., *Enumerability. Decidability. Computability*, Springer, 1965.
Hopcroft, I. E. + Ullman, J. D., *Formal Languages and their relation to automata*, Addison-Wesley, 1969.
Kleene, S. C., *Introduction to Metamathematics*, North-Holland, 1971.
Salomaa, A., *Formal Languages*, Academic Press, 1973.
Smullyan, R. M., *Theory of Formal Systems*, Princeton University Press, 1961.

E. García Camarero

L'ÉVOLUTION DU LEXIQUE CASTILLAN

Bien que l'espagnol possède une très belle História de la lengua española de R. Lapesa —la meilleure de toutes les langues ibéroromanes— on trouve très souvent des lacunes dans des études plus spéciales et détaillées. P. ex. dans le *Manual de Lingüística Románica* (I. Iordan - M. Manoliu, II, 148) les auteurs prétendent que l'enrichissement des langues romanes se soit fait surtout par latinismes et héllénismes, et cela à partir de 1800. Avec la partie castillane de notre *Dictionnaire chronologique des langues ibéroromanes* il sera facile de vérifier ces constatations. Nous avons réuni 23.588 mots espagnols dont 14.103 (59 %) portent une date de première apparition et dont 9.485 n'en portent