

## EL LENGUAJE Y LAS MATEMÁTICAS

Tal vez parezca atrevimiento, y seguramente lo sea, que alguien dedicado a los estudios sobre el hombre meta sus narices en el huerto ajeno. ¿Puede enseñar algo el lingüista al matemático? ¿Y el matemático al lingüista? ¿Tienen ambas ciencias alguna parcela en común? Vaya por delante que no le pidan peras al olmo ni un número racional a la raíz cuadrada de dos. Yo sólo conservo residuos de mi viejo oficio de humanista y ni siquiera los primeros elementos de una disciplina de la cual ignoro casi todo.

He tomado como tema de esta conferencia un punto de encuentro entre el saber de las letras y el saber de los números. Hablo de la conocida ley de Zipf. La enunciación, sin señalar todo el aparato técnico de los especialistas en probabilidad, es muy sencilla: en un texto de cierta longitud – por ejemplo éste - la segunda palabra que más se repite lo hace la mitad de la primera; la tercera palabra más repetida es el tercio de la primera. Y así sucesivamente. ¿No es esto un hecho sorprendente y, al mismo tiempo, inquietante? Si nos desconcierta es porque sospechamos que, de alguna manera, nuestro lenguaje y, por tanto, nuestro

pensamiento, se halla determinado por ciertas leyes ajenas a nuestra voluntad. La hipótesis de Laplace haría una brecha en el último reducto de la libertad humana. ¿Y cuál será todavía más nuestro estupor oyendo decir que esta ley se cumple estrictamente en todas las lenguas habidas y por haber, ya sea el inglés o la lengua bantú, el sánscrito, el hitita desaparecido, el persa antiguo, las lenguas esquimales y, además, incluso en lenguas artificiales como el esperanto? ¿Y eso que de las cinco mil lenguas de cuya existencia sabemos, los lingüistas están todavía muy lejos de conocer sus gramáticas! Cuenta Camilo José Cela la historia de un niño que entra en un casino acompañado de su padre. “Vamos a ver – le dice al niño – ¿cuánto es cinco millones doscientos treinta y siete mil noventa multiplicado por un millón seiscientos mil cuarenta y cinco? Y el niño, más en un santiamén que en un avemaría, responde ante la admiración de todos los presentes una cantidad elevadísima. Por supuesto, todos se maravillan del talento del niño salvo una persona que saca del bolsillo su calculadora y dice: “no es cierto.” Pero, aunque no se trate de ningún silogismo, concedamos la mayor, pues las calculadoras no engañan ni exigen un acto de fe como los milagros. Doctores tiene la Iglesia y Doctores tiene también la informática.

Ahora bien, hemos de decir que se trata aquí de una ley

empírica, “a posteriori”, y no el resultado de una deducción lógica. O sea, “a priori”. Podemos realizar un experimento verbal: tomemos un diccionario eligiendo doce mil veces una misma palabra, seis mil veces otra, cuatro mil veces la siguiente, etc. Y hagamos que su distribución cumpliera la ley de Zipf. Pero ¿expresarían los mismos mensajes? Cualquiera puede ver que no pues nada se dice sobre cuáles sean las palabras utilizadas. En un libro de poemas sería extraño que apareciera la voz “ornitorrinco” y un libro de zoología no usaría mucho la palabra “amor”. Bastante complicado sería, además, barajarlas. Y ni siquiera podría escribirse nada si dichas palabras fuesen incompatibles para formar una frase sintácticamente correcta. ¿No podemos tomar diez mil veces la misma conjunción? En suma, un escritor no podría escribir nada si tuviese de antemano ya tales palabras dadas en las condiciones establecidas y estas palabras no pudiesen combinarse para crear un texto coherente. Se invierte la relación. Primero crear el texto, luego comprobar empíricamente la relación matemática y no al revés.

Esta ley misteriosa se basa exclusivamente en criterios formales y no semánticos. ¿Y qué se cuenta? Pues las palabras tomadas como unidades discretas, separadas por el espacio vacío, o blanco, que las separa en la escritura. En realidad esto no

sucede así pues una cosa es la letra y otra distinta el lenguaje verbal. Entre las palabras hay un flujo sonoro, continuo, y solamente el análisis gramatical – no todas las categorías tradicionales son aplicables a todas las lenguas - permite hacerlas visibles. La lectura se realiza con el ojo, no por el oído. Un niño, ignorante de la gramática, tendría dificultades para escribir “la casa” separando el artículo del sustantivo. En esa cuenta de las palabras, para cumplirse la ley de Zipf, no importa el orden de dichas palabras y el cual les da un sentido preciso. Es lo mismo decir “pobre hombre” que “hombre pobre” y tanto da “el coche azul” como “coche el azul”. A la hora del cómputo ello es impropio.

Pues bien, hemos dicho que para contar palabras éstas han de ser unidades aisladas, distinguidas por el vacío o blanco que las separa. Sin el corte del espaciador ¿cómo podría distinguir un ordenador el verbo “escoja” de “es coja”, o sea, con piernas de distinto tamaño. El ordenador precisa que le den órdenes. El problema es que los lingüistas no han llegado a ponerse de acuerdo en una definición universal de la palabra que sea válida para todos los idiomas del mundo. ¿Cómo? ¿Si ello se ve tan claro? Ya dijimos que un niño escribe “lacasa” todo junto pues el artículo femenino singular está fusionado de una manera inseparable,

fonética y no gramaticalmente, al nombre. ¿Y qué ocurre con las lenguas aglutinantes? ¿Es la palabra la agregación de los monemas independientes? ¿O bien cada monema separado? El origen del verbo futuro “cantaré” es la perífrasis del infinitivo con el auxiliar haber: “cantar he”, “he de cantar”. Pero hoy están fusionados los dos verbos en una sola forma. Pongamos también un caso esclarecedor. Parafraseando a Jorge Manrique, no hablaremos de los tirios y troyanos cuyas hazañas no vimos. No vamos a hablar de las lenguas africanas y amerindias, sino de la cercana lengua francesa. Si digo la frase “je ne sais pas”, la gramática nos dice que “je”, pronombre, es una palabra. Pero en francés, fuera de la interposición, el pronombre *je* es inseparable del verbo. En español podemos decir: “Hablo” o bien “yo hablo”. En francés “je dorm, je cours”, el pronombre *je*, aunque escrito separado, forma cuerpo con el verbo. Y entonces, a la hora de contar, en la frase “je dorm, je cours” ¿cuántas palabras hay? Podemos contar tres palabras distintas, dos veces *je* y una vez *dors* y *cours*, o bien dos palabras “jedorm”, “jecours”. Esto sería semejante a los prefijos de los vocablos “teléfono” y “polideportivo” donde “tele” y “poli” no son independientes. En español “don Juan” y “don Nadie”. ¿Es una palabra? ¿O dos palabras juntas como “sacacorchos”? Hoy tomamos como una palabra “etcétera” cuando

en latín son dos: “et” y “caetera”. Las normas actuales podrían retroceder al estado inicial. Claro está que para el matemático, atento únicamente a un criterio formal, esto es indiferente. Solamente le interesa que haya una secuencia, una sucesión de formas blancas que delimiten otras formas negras y las cuales se cuentan repetidas cada una un cierto número de veces porque son iguales. No importa lo que signifiquen, carne, pescado o alcornoque.

Acabamos de decir que al matemático no le importa el sentido. Solamente cuenta y establece una relación. Para él lo mismo le da decir “Pedro tiene dolor de cabeza” que “tiene de dolor cabeza Pedro”. Vamos a suponer que las palabras de un vocabulario normal son sustituidas por fichas de números del 1 al 20.000 y que cada ficha puede usarse tantas veces como se quiera. Se supone que al distribuir las al azar debería cumplirse la ley de Zipf. Así, pues, el misterio habría salido del ámbito exclusivo del lenguaje para situarse en el terreno interior de la matemática. Pero ¿esto es así? ¿No estaremos haciendo trampas echando en el cubo del azar unos dados cargados? Veamos: en un discurso no todas las palabras tienen la misma probabilidad de aparecer. Yo puedo decir: “el libro”, “el perro”, “el coche”. No puedo invertir el orden como se hace en la lengua rumana. El artículo “el”, que se

repite tres veces, acompaña necesariamente a los tres distintos nombres. Pero sustantivos hay muchos y artículos hay pocos. Tenemos que el número total de artículos masculinos en singular, siempre el mismo, es igual al número total de sustantivos masculinos en singular. Aquí, en cambio, distintos, cada perro de su pelaje. El número de preposiciones, invariables, está sobre la veintena; el número de conjunciones, también inalterables, tampoco es alto. Los artículos, repartidos según género y número, reúnen a unas cuantas formas. Y lo mismo se podría decir de los adverbios. En resumen, habas contadas. Estas palabras, para decirlo de algún modo, son “estructurales”. Privilegiadas. Ellas forman el cañamazo sobre el cual se tejen los variados bordados. Por el contrario, sustantivos y verbos son abundantísimos en un vocabulario de miles de palabras. Quedan diluidos. Y como son muchos, sus probabilidades de aparecer en el discurso también son mucho más limitadas. Y a esto se añade la lógica o, si queremos, la restricción semántica. Podemos decir: “el caballo teje los libros mientras que los caracoles leen las farolas”. La estructura de esta frase es absolutamente correcta conforme a la gramática. En cambio no se dice “caballo el salta”. No es un orden correcto. La poesía surrealista se apoyó en esta posibilidad de la lengua de conservar la sintaxis atentando contra la semántica. El

lenguaje permite decir muchas más cosas de las que se puede entender. No todo lo racional -como dice Hegel- es real y todo lo real es racional.

Volvamos a las posibilidades de aparecer las palabras en un discurso. Si digo “el (vacío) me ha mordido”, en el hueco pueden darse tres casos. “el libro me ha mordido” es imposible”, “el jefe me ha mordido” es bastante improbable; “el perro me ha mordido es bastante creíble”. En suma, las tres palabras – libro, jefe, perro – tienen distinto grado de probabilidad de aparecer en la cadena del discurso. Por tanto el cómputo del número de palabras que aparecen en un texto no es igualitario, no todas las palabras tienen la misma oportunidad. Pensemos en el comienzo del Quijote: “En un lugar de la Mancha”. Si Cervantes quiere comenzar la frase tiene tantas posibilidades como palabras existen. Podría decir: “Pedro”, “Madrid”, etc. Cualquiera. Esto es el terror a la “página en blanco” que asusta al escritor en espera de que descendan las musas por la tramoya. Ahora bien, una vez que ha escrito “En un lugar de la Mancha”, entre la palabra “En” y la palabra “lugar” solamente serían posibles voces como “el”, “este”, “ese”, “mi”, y alguna más. De manera que mientras la escritura está en acto el escritor se halla libre, pues aunque las palabras “estructurales” - el artículo es el artículo -, le constriñen,



además de la lógica, la abundancia de nombres y verbos entre los cuales elegir le da suficiente holgura como para escribir desde “Buenos días” hasta la “Crítica de la razón pura” de Kant. ¿Qué quiere esto decir? Pues sencillamente que mientras el discurso está en curso no puede hablarse de la ley de Zipf. El texto debe estar completo, cerrado del todo. Es evidente, pues ¿cómo se pueden contar las palabras si éstas aumentan constantemente? ¿No se debe esperar al final del recuento de votos para conocer el ganador? En una canción se dice “Buenas noches, señora, recuerdos a su señor”. Pero si el autor corrige y repite un segundo “buenas noches, señora”, no solamente cambia el sentido (ahora se insinúa un adulterio) sino que también se ha alterado el cómputo de las palabras.

Así, pues, se podría suponer que las palabras más frecuentes en el texto en un orden decreciente son aquellas más posibles de figurar en el contexto, las que tienen mayor certeza y, por tanto, aportan menor información, un principio muy elemental, como es de sobras conocido por todos. Pero esto no es claro, pues una palabra que tiene muchas probabilidades de salir en un contexto parcial puede tener muy pocas posibilidades de salir en el contexto total.

La ley de Zipf establecía una correspondencia fundada en

un criterio exclusivamente formal. En lugar de letras se hubiesen podido emplear matices de colores para cada una. Sin embargo, desde esa ausencia de sentido, hemos vuelto de nuevo al significado del lenguaje normal. La semántica que habíamos dejado en la puerta se cuelga de nuevo por la ventana. Sólo una mente concedora de los significados puede llenar los huecos de una frase. Nadie, siguiendo un criterio exclusivamente formal, podría hacerlo en un idioma que no sabe. A lo sumo, en castellano, sabemos que las palabras siempre acabarán en vocal, n, s o líquida, l y r (lo cual, aunque esto sea harina de otro costal, justifica las normas de acentuación).

Dentro de un contexto dado, hemos visto, hay palabras imposibles, improbables y probables. Hay unos órdenes correctos de las oraciones, con mayor o menos libertad o rigidez, y otros órdenes que violan la gramática. En una determinada posición solamente pueden aparecer unas determinadas palabras. Cada palabra, tomada individualmente, se opone lexicalmente a todo el resto de las otras palabras. Sin embargo, no se opone a todas ellas en el sistema de la lengua, pues el artículo “el” se opone a “perro” o “día” aunque no lo hace con la conjunción “y” detrás, pues esa estructura está prohibida. El discurso vendría a ser como una sucesión de palabras con un rango de probabilidad variable entre

el cero y el uno. Hasta una palabra idéntica, dijimos, no tiene la misma probabilidad de aparecer en distintas posiciones pues sus vecinas palabras no son las mismas. Si se pudiese reflejar en una gráfica los valores de probabilidad de cada palabra en el discurso dibujarían una línea ondulante irregular. Pienso, en mi ignorancia sobre los conceptos de la estadística y su terminología, que la probabilidad de cada palabra en un texto sería la relación entre el número de palabras que pueden ocupar ese “hueco” y el conjunto total de palabras del vocabulario.

Yo no sé si lo que he dicho tiene mucho sentido. No lo sé. Como ya he dicho en un comienzo, estoy muy lejos de poseer una formación matemática.

Pero sí quiero confesar que, después de la fascinación que me produce una ley tan misteriosa, uno alberga el deseo oculto de que sea falsa, de que tenga un talón de Aquiles. O, al menos, sea tan comprensible como el hecho de que cada lengua tenga un porcentaje aproximado de aparición de los fonemas – fundamento de la criptografía - y que está en función de su estructura fonológica. Claro está que una ley empírica solamente es falsa si los datos que se manejan son también falsos. Deseo que sea errónea. Aunque sea para sentir la satisfacción de que mis palabras no están sujetas a ninguna norma externa. Ahora bien,

ya dije antes que el escritor es libre mientras escribe aunque una vez escrito el texto parezca determinado y sujeto al yugo de la ley de Zipf. Aquí podríamos hacer un parangón saltando a la teología. El hombre tiene libre albedrío para actuar aunque Dios, omnisciente, sepa simultáneamente cómo va a actuar. Pero no nos perdamos en controversias pelagianas. Aceptemos que esta ley mágica existe en unas cuantas lenguas. Ello bastaría para darle su halo mágico (creer que afecta a todas las lengua es convertir las sardinas en ballenas).

Aunque, como ya he señalado reiteradas veces, no soy matemático, ni soy tampoco demógrafo, tengo algunos ligeros conocimientos sobre historia política. Siempre hay gente que necesita descubrimientos excepcionales. ¡Se ha encontrado cura al cáncer!, exclama algún periódico o, incluso, alguna revista científica divulgativa. Pues bien, llevados del entusiasmo se ha llegado a afirmar que la ley de Zipf se cumple también en el tamaño de las ciudades. ¿Es aplicable a todas como en el caso de las lenguas? La historia nos demuestra que los Estados cambian sus fronteras debido a guerras, tratados internacionales, etc. Decía Unamuno que hablaba de sí porque era el hombre que tenía más cercano. Yo, como aragonés, puedo decir que la enorme distancia entre Zaragoza y la segunda población se debe

solamente a circunstancias históricas de naturaleza económica. Y si Cataluña se independizase como República, Barcelona no sería el doble que Manresa, Sabadell o la que fuese. Y si España permanece unida, ¿no debe aumentar mucho Madrid para doblar a Barcelona? Se han hecho estudios sobre las ciudades de Estados Unidos, pero es evidente que si en la guerra de secesión el sur hubiese vencido las condiciones en dos países diferentes también serían distintas. Y de la misma manera que nos preguntábamos si en francés “je cours” es una o dos palabras, podemos interrogarnos si estando inseparables Burjasot, Burjasot-Godella y Godella son una o tres poblaciones. ¿Cuál es el cómputo correcto? ¿Cómo realizarlo? Si las tomamos como una, ¿da el mismo resultado que si las tomamos como tres? Y si no podíamos realizar la ley de Zipf en un texto que no fuese cerrado, pues las palabras contadas aumentan, ¿han dejado de crecer las ciudades? Y si tienden a cumplir con su crecimiento dicha ley, el bombardeo de una ciudad ¿no puede derribarla como hace con las casas?

Para terminar ya, probablemente tantas veces oír probabilidad en la boca de una aficionado no les haya a ustedes probado nada. Si no pruebas, espero al menos merecer su aprobado habiéndoles sembrado algunas dudas sobre la misteriosa ley de Zipf. Y la duda es el inicio de la ciencia.

Pablo Galindo Arlés

29 de octubre de 2019