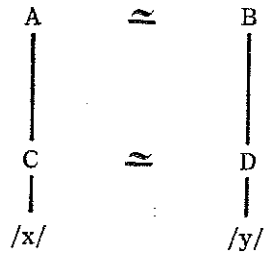
PROBABILIDAD Y DETERMINACIÓN ETIMOLÓGICA

1. En una Conferencia dedicada a la Lingüística Aplicada que se celebró en 1960, en la ciudad de Cernăuţi (Chernovitsy), en la República Soviética de Moldavia, N. F. Pelévina presentó una comunicación sobre un particular cálculo probabilístico destinado a determinar —a falta de argumentos filológicos pertinentes— si, en una lengua y en unas condiciones dadas, un grupo de palabras fonológicamente afines o idénticas deben su parecido a un azar o a un origen etimológico común [1]. El método es bien simple y su justificación parece tanto mayor cuanto más escasos sean la información histórica y los antecedentes documentales al respecto.

Trataremos, en lo que sigue, de examinar los rasgos principales de esta propuesta, toda vez que no ha merecido, por lo que sabemos, una atención extensa entre los especialistas.

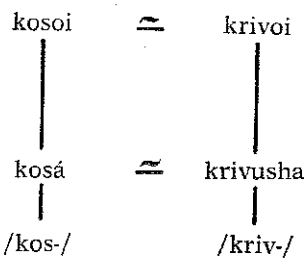
2. El procedimiento enuncia que cuanto más pequeño es el valor del producto de las probabilidades de aparición de un número de elementos léxicos con una estructura fonológica y semántica al menos parcialmente análoga, tanto mayor es la certeza de que dichos elementos léxicos procedan de una misma base histórica. Si el valor de aquel producto tiende a cero, la unicidad originaria de los elementos implicados tiende a infinito, y viceversa. Más en detalle, considérense en una lengua dada dos o más grupos de palabras con una determinada afinidad semántica, pongamos *A-B*, por un lado, y *C-D*, por otro:

[1] N. F. Pelévina, «Ustanovlenie etimologuicheskogo tozhdestva s pomoshchu umnozhenia veroiatnostei» [Determinación de las identidades etimológicas con ayuda de la multiplicación de probabilidades], *Pitania Prikladnoi Lingvistik*, Cernăuţi, 1960, págs. 38-40. Cf. también en S. Marcus, Ed. Nicolau y S. Stati, *Introducere în lingvistica matematică* [Introducción a la lingüística matemática], Bucarest, 1966, págs. 110-2. Para la transliteración del ruso seguimos las normas de J. Calonge en *Transcripción del ruso al español*, Madrid, 1969.



Si *A* y *C* tienen en común una estructura fonológica /x/, mientras *B* y *D* tienen otra en común /y/, la probabilidad de la relación semántica (a ser posible, de sinonimia) entre *A* y *B* equivale al producto de las probabilidades de las secuencias /x/ e /y/. Y como el mismo valor vale también para la relación semántica entre *C* y *D*, hay que multiplicar aquel producto por sí mismo. Si el valor resultante es tan pequeño que tiende a cero, es consecuentemente probable que *A* y *C*, por una parte, y *B* y *D*, por otra, hayan tenido un origen etimológico idéntico.

Para ilustrar el principio expuesto, Pelévina se sirvió de ejemplos del ruso, si bien ya conocidos en sus antecedentes históricos. Veámoslo en una aplicación del esquema anterior.



Comparó las palabras *kosoi* [косой], 'oblicuo', y *kosá* [коса], 'guadaña', aquí presuntamente emparentadas, y extrajo los elementos fonológicos idénticos: /kos-/. Estimando en 50.000 las palabras de un vocabulario de la lengua rusa y en 100 las que empiezan por la secuencia /kos-/, la probabilidad de que una palabra rusa empiece por /kos-/ es, según esto, de 100/50.000, es decir, 1/500. Ahora bien, el adjetivo *kosoi* es quasi-sinónimo de *krivoi* [кривой], 'torcido', y, además, *kosá* lo es de la palabra dialectal *krivusha* [кривуша] [2].

[2] Las acepciones más usuales de estos cuatro términos pueden distribuirse así: *kosoi*, 'oblicuo, torcido, inclinado' // 'bizco' // 'liebre'; *kosá*, 'guadaña'

Esta otra pareja léxica ofrece, de nuevo, una secuencia fonológica común formada por /kriv-/; y como el número de palabras rusas que empiezan por /kriv-/ llega a unas 50, su probabilidad oscila en torno a 50/50.000, o sea a 1/1.000. Es decir, que admitiendo por simple observación la sinonimia o quasi-sinonimia de *kosoi* y *krivoi*, *A* y *B*, por una parte, y la de *kosá* y *krivusha*, *C* y *D*, por otra, resulta que tanto *A* y *C* (*kosoi* y *kosá*) como *B* y *D* (*krivoi* y *krivusha*) manifiestan una parcial identidad fonológica. El procedimiento consiste, pues, en averiguar si esta conjeturable interrelación alterna de significados y significantes que presentan los cuatro términos obedece a una mera casualidad o deriva de un efectivo parentesco etimológico. Y, en efecto, como la sinonimia manifiesta de *kosoi* y *krivoi* se establece entre una palabra con la secuencia /kos-/ y otra con la secuencia /kriv-/, es decir entre dos probabilidades de 1/500 y de 1/1.000 respectivamente, el valor de la probabilidad sinonímica de *kosoi* y *krivoi* equivale al producto

$$\frac{1}{500} \cdot \frac{1}{1.000} = \frac{1}{500.000}$$

Pero como las probabilidades de /kos-/ y de /kriv-/ afectan igualmente a la otra pareja de sinónimos, *kosa* y *krivusha*, ocurre entonces que las mismas secuencias fonológicas se emplean dos veces para expresar un significado único, de modo que la doble sinonimia resultante tiene una probabilidad de

$$\frac{1}{(500.000)^2} = \frac{1}{250.000.000.000}$$

valor que, evidentemente, tiende a cero. La conclusión dice que la ínfima probabilidad conjunta de *kosoi*, *kosá*, *krivoi* y *krivusha* no se debe a un azar, sino a una misma base etimológica para *kosoi* y *kosá*, por un lado, y para *krivoi* y *krivusha*, por el otro, tal como lo confirma, por cierto, la historia del ruso.

3. Para examinar los múltiples aspectos de esta propuesta, es preciso subrayar de antemano el carácter tensivo —ya que no ten-

// 'trenza'; *krivoi*, 'curvo, torcido, tortuoso, oblicuo' // 'tuerto' // 'falso'; *krivusha*, aparte del uso dialectal consignado por Pelévina, constituye, en su lectura más inmediata, un diminutivo de *krivoi* ('tuerto').

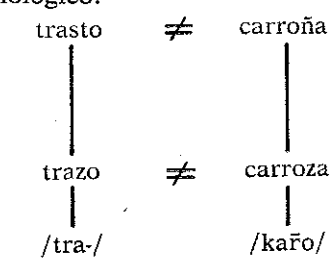
dencioso— de todos los procedimientos estadísticos y, naturalmente, de éste. La misma Pelévina insistió en que los hechos sólo propenden a comportarse así y que ello no se advierte sin considerar un número elevado de eventos.

Una formulación más intuitiva (aunque, como veremos, muy incompleta) del principio expuesto denuncia, queramos que no, una lógica inmediata. A lo sumo, nos dice que si en una lengua natural, con millares de estructuras fonológicas diversas, hay una determinada estructura que sólo aparece en dos palabras, es muy probable, pero no obligatorio, que esta coincidencia insólita no se deba al azar, sino a un vínculo semántico, más o menos persistente desde el pasado, entre aquellas dos palabras (tal es el caso de *israelí-israelita, istmo-ístmico,* etc.). La originalidad principal del método de Pelévina estriba en el hecho de ofrecer una vía estadística que permite matizar los grados de probabilidad para cuando no se cumple esta condición tan extrema. Justamente por ello y para reducir al mínimo los efectos de la casualidad homonímica, introduce una condición basada en el cotejo de palabras por parejas (o por grupos superiores) junto con la intervención de algunos supuestos semánticos. Nótese que esto no contiene ninguna contradicción, ya que presupone el reconocimiento de unas analogías fonológicas alternas, pero no correlativas, respecto de otras analogías semánticas; es decir, se establece que si hay una analogía semántica entre *A-B* y *C-D* al mismo tiempo que una analogía fonológica entre *A-C* y *B-D*, la probabilidad de que *A-C* y *B-D* sean etimológicamente sinónimas depende de la peculiaridad, dentro de la lengua en cuestión, de su analogía fonológica conjunta. Más en general, en unas condiciones dadas, cuanto mayor sea el número de grupos léxicos así implicados tanto mayor será, en consecuencia, el grado de peculiaridad analógica y, por tanto, de convergencia en la etimología.

Naturalmente, el requisito del agrupamiento múltiple hace aumentar la fiabilidad del método, pero restringe en igual medida su capacidad de aplicación. Es, por lo demás, necesario, puesto que, de otro modo, estructuras fonológicas muy aisladas, tales como en castellano /aθt-/, en *azteca* y *aztor*, /kθar-/, en *czar* y *czarda*, y muchas otras más complejas productoras de homonimias, como en *bote, limón, escatología*, etc., entrarían en conflicto con el supuesto referido, pese a que, por ejemplo, la probabilidad de la sinonimia de *escatolo-*

gía$_1$ (< ἔσχατις, 'extremo, último') y *escatología*$_2$ (< σκῶρ, σκατός, 'excremento') —con una coincidencia fonológica absolutamente peculiar— alcanzaría un resultado imposible de refutar al compararlo con cualquier otro grupo afín. La incongruencia de este eventual resultado queda, entonces, teóricamente asegurada por la enorme dificultad de encontrarles a *escatología*$_1$ y a *escatología*$_2$ al menos dos sinónimos, *B* y *D* respectivamente, tales que presenten a su vez un elevado grado de analogía fonológica [3]. Sea como sea, si se establece que las palabras *A* y *C*, fonológicamente próximas, han de ofrecer, respectivamente, otras palabras *B* y *D*, sinónimas o quasi-sinónimas, que resulten igualmente próximas desde el punto de vista fonológico, para someterlas, sólo entonces, al examen probabilístico, hay que admitir que tales condiciones —aun en su manifestación más simple— no abundan en las lenguas naturales. Una prueba inmediata la tenemos en el hecho mismo de que Pelévina hubo de recurrir a un uso dialectal en su ejemplo.

Habría que aclarar o indagar, en todo caso, el grado y el tipo de relación semántica inicial requeridos a fin de considerar pertinentes los resultados de la comparación (¿hasta qué punto deben ser sinónimos *A-B* y *C-D?)*. La importancia de esta cuestión queda fuera de toda duda y su necesidad es de todo punto decisiva, porque si la determinación de la sinonimia etimológica se hiciera meramente a través del propio cálculo probabilístico y al margen de todo reconocimiento previo de analogía semántica entre los términos alternos, entonces podría incurrirse en verdaderos espejismos. Efectivamente, basta analizar dos pares léxicos fonológicamente afines pero sin relación semántica alguna, como *trasto-trazo* y *carroña-carroza*, para comprobar que los resultados de sus probabilidades nos llevan a un fuerte equívoco etimológico:



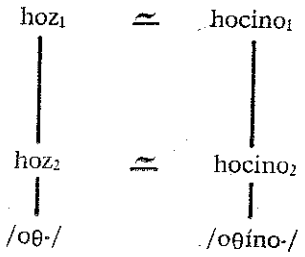
[3] Por supuesto, nociones tales como cruce de palabras, préstamo, cultismo y otras que resulten directamente de comprobaciones diacrónicas previas no pueden aquí aducirse sin contradicción con el método probabilístico.

El primer par tiene en común la secuencia fonológica /tra-/ y el segundo /kaɾo-/. Si estimamos también en 50.000 las palabras de un vocabulario castellano, en unas 700 las que empiezan por /tra-/ y en unas 16 las que empiezan por /kaɾo-/, con una probabilidad de 1/72 y de 1/3.125, respectivamente, la probabilidad de la sinonimia de *trasto* y *carroza* junto con la de *trazo y carroza* equivale a:

$$\frac{1}{(72 \cdot 3\ 125)^2} = \frac{1}{50.625.000.000},$$

lo que no está nada mal, habida cuenta que entre estas cuatro palabras no existe ningún vínculo etimológico.

La búsqueda, entonces, de grupos léxicos más o menos sinónimos y más o menos homónimos en forma alternativa es muy peliaguda, y aun así tampoco faltan casos engañosos. Tal sucede, por ejemplo, con los homónimos *hoz (< falce / fauce)* y sus quasi-sinónimos *hocino*:



En una supuesta ausencia de pruebas históricas, tratemos de averiguar si *hoz* $_1$, 'instrumento para segar', y *hoz* $_2$, 'estrechura de un valle o un río en lugar profundo', con una estructura fonológica idéntica /oθ-/, tienen un origen etimológico común o no. Ambos ofrecen dos palabras quasi-sinónimas e igualmente idénticas desde el punto de vista fonológico; respectivamente, *hocino* $_1$, 'instrumento para cortar leña', y *hocino* $_2$, 'angostura de un río entre montañas', con una secuencia común /oθíno/. Si continuamos estimando en 50.000 las palabras de un vocabulario castellano, la probabilidad de aparición de la secuencia /oθ-/ equivale, en números redondos, a 1/1.500 (sobre unas 30 palabras), mientras que la de /oθíno-/ alcanza un 1/25.000 (sólo aparece en las dos palabras cotejadas). Con ello, la probabilidad sinonímica de *hoz* $_1$ y *hocino* $_1$ depende, como se ha

visto, de la probabilidad conjunta de una secuencia fonológica /oθ-/ y de otra /oθíno-/, que se obtiene multiplicando los valores anteriores:

$$\frac{1}{1.500} \cdot \frac{1}{25.000} = \frac{1}{37.500.000},$$

resultado que debe coincidir con el de la probabilidad sinonímica de *hoz* $_2$ y *hocino* $_2$. Entonces, para calcular la probabilidad sinonímica conjunta de las parejas respectivas, hay que tener en cuenta las probabilidades de las secuencias /oθ-/ y /oθíno-/ dos veces, o sea:

$$\frac{1}{(37.500.000)^2} = \frac{1}{1.406.250.000.000.000},$$

cifra muchísimo más tendente a cero que la propia ilustración de Pelévina. Y, sin embargo, las pruebas filológicas tampoco corroboran aquí la conclusión de que *hoz* $_1$ y *hoz* $_2$, por un lado, y *hocino* $_1$ y *hocino* $_2$, por otro, «deban» tener una base etimológica común.

Faltaría investigar, en todo caso, si nos hallamos ante un simple contraejemplo o si, por el contrario, es el ejemplo de Pelévina producto de una coincidencia fortuita. Habría, en suma, que multiplicar las observaciones al máximo para dar una contestación cabal a esto. El problema principal es, sin duda, de tipo práctico, tal como hemos venido reiterando; y es que, ¿cuántos ejemplos de sinonimia y homonimia alternas existen en las lenguas naturales? Tal vez la respuesta tenga que ver con la estructura tipológica de cada lengua. A nosotros, el rastreo afanoso de casos por páginas y más páginas del diccionario nos ha conducido a un solo ejemplo, formalmente óptimo (veinte veces mejor, a la vista de los resultados, que el propuesto por Pelévina), pero fallido. Casualmente fallido, a nuestro juicio, porque el método en cuestión es teóricamente válido: dos palabras etimológicamente sinónimas serán, o tenderán a ser, más semejantes desde el punto de vista fonológico que otras dos palabras cualesquiera que no mantengan ninguna relación histórica recíproca. Obsérvese que este principio vale también para todas las relaciones

derivativas entre términos léxicos con o sin cambio en la categoría gramatical respectiva (cf. *escribir-escritura-escriba-escritor*, etc.)[4].

Pero, de nuevo, su aplicación a la realidad presenta otros importantes cabos sueltos en su estrategia metodológica inmediata. Y es que habría que precisar también el uso exacto de la expresión «semejanza fonológica», ya que aquí parece estar sujeto no sólo a la estructura actual, sino también a otras pretéritas, de una misma lengua. De otro modo, cabría el riesgo de considerar más afines, digamos, *pie* y *piedra* que *pie* y *pedicuro*, para mencionar un solo caso entre muchísimos otros posibles. En general, la gran mayoría de dobletes —así como analogías diacrónicas, etimologías populares, etcétera— pasaría inadvertida ante una prueba escuetamente probabilística[5]. En cambio, la comparación de adverbios en *-mente* a menudo induciría a involucrar raíces diversas en un cotejo sin sentido. Piénsese, por ejemplo, en dos adverbios como *vilmente* y *hábilmente*, con una larga secuencia /-bilménte/ en común. Esto plantea asimismo el problema de decidir qué sección —inicial, media o final— o qué secciones de la palabra serían objeto de la comparación fonológica a fin de no tratar de igual modo, pongamos, *amor*, *humor* y *humo*. Para cada lengua, pues, habría que sentar de antemano unas referencias concretas de tipo fonológico e incluso ortográfico (en castellano, distinción entre *b* y *v*, presencia o ausencia de *h* y otros datos que fuesen relevantes) antes de poner en marcha la comparación[6].

---

[4] Eso al margen de las profundas alternancias fonológicas incluso en el ámbito de los paradigmas nominales y verbales, a menudo irregulares, en cualquier lengua flexiva. A su vez, el descubrimiento de relaciones derivacionales en la lengua tomada como referencia etimológica resulta enormemente precario, a la luz de los resultados modernos (cf. *ocho* y *octavo*, *meter* y *premisa*, etc.).

[5] Tras la lectura de este trabajo (VI Simposio anual de la Sociedad Española de Lingüística, Madrid, diciembre de 1976), el Prof. Sebastián Mariner formuló la interesante sugerencia de utilizar este método probabilístico para explicar los casos de etimología popular. A nuestro juicio, el fenómeno en cuestión puede ser efectivamente descrito (y acaso predicho) a través de un modelo basado en este tipo de cálculo, pero no del todo igual al presente. Confiamos, para un próximo futuro, ofrecer una elaboración sobre las incidencias principales que aparecen en esta particular modelación.

[6] De no observar las diferencias ortográficas, el peligro de las falsas adscripciones etimológicas no haría más que crecer, al menos en la mayoría de los casos. Así ocurriría, por ejemplo, con las palabras *ingerir* (< *ingerĕre*) *e injerir* (< *inserĕre*), con una secuencia fonológica totalmente idéntica, /inxerír/, y con dos quasi-sinónimos, *zampar*$_1$, 'comer (apresurada, descompuesta y excesivamente)', y *zampar*$_2$, 'meter una cosa en otra (de prisa y de suerte que no se

Digamos, mientras tanto, que este cometido no sería nada fácil de llevar a cabo y que contendría, con seguridad, numerosos círculos viciosos con respecto al propósito mismo que persigue el método.

Otro inconveniente serio radica en la interpretación de los datos numéricos. En rigor, se plantea una cuestión que no admite términos medios: la de decidir si un grupo de palabras tiene un origen etimológico idéntico o no. Los resultados, en cambio, parecen sugerir una cantidad innecesariamente grande en grados de parentesco histórico, que no cabe utilizar para afirmar siquiera que en el ejemplo castellano de *hoz* y *hocino* hay una certeza más de veinte veces mayor que en el ejemplo ruso del principio, aun prescindiendo de lo que dice al respecto la filología fundamentada. Hay, en suma, que contestar a la pregunta, crucial, de cuándo o a partir de cuándo se puede considerar una probabilidad suficientemente pequeña a fin de que el examen adquiera valor probatorio. Este requisito lleva además implícita la necesidad de buscar unos criterios coherentes para determinar el número total de palabras homologadas en un vocabulario e incluso para reconocer previamente si un vocabulario es el marco más adecuado para calcular la probabilidad de aparición de una o más palabras. Como se sabe, los índices de probabilidad varían de un modo radical si la comprobación se efectúa sobre ocurrencias extensas, escritas u orales.

Como sea que estadísticamente el fonema /a/ es el más frecuente en castellano, digamos que el cotejo de cuatro palabras del tipo

---

vea)', igualmente idénticos desde el punto de vista fonológico: /θampár/. Como las respectivas secuencias sólo aparecen en las cuatro palabras consignadas, la probabilidad total de esta sinonimia, sobre 50.000, equivale a

$$\frac{1}{(25{,}000)^4} = \frac{1}{390.625.000.000.000.000}$$

valor que, por cierto, es el mínimo posible que cabe obtener en una comprobación a base de dos parejas léxicas. Como veremos a continuación, todas las probabilidades del castellano (siempre sobre 50.000 palabras) oscilarían entre un máximo de 1/2.401 y este valor, más de un millón y medio de veces inferior al que obtuvo Pelévina. Pero el mero cruce que ha existido entre *ingerir* e *injerir* nada tiene que ver con el resultado, aunque al menos éste siente la identidad etimológica de *zampar*$_1$ y *zampar*$_2$, cosa que nadie ha puesto en duda aun sin testimonios documentales anteriores al siglo XVII (cf. J. Corominas, *Diccionario crítico etimológico de la lengua castellana*, Madrid, 3.ª r., 1976, volumen IV, s. u.).

abeja, apícola, año y *anual,* con una común estructura fonológica mínima, /a-/, y sin ninguna relación semántica entre *abeja-año* y *apícola-anual,* daría una probabilidad de sinonimia conjunta, sobre unas 7.200 palabras empezadas por /a-/, inversamente proporcional a algo menos de

$$\frac{1}{(7)^4} = \frac{1}{2.401}$$

Este es, en teoría, el valor máximo que puede alcanzar en castellano la probabilidad sinonímica de cuatro palabras a la vez (aunque en nuestro ejemplo hayamos deslizado dos emparejamientos etimológicos). Los valores de cualquier otra comprobación serán iguales o inferiores a éste, de modo que no podemos decir, entonces, que 1/2.401, es decir, 0,000.416.4 (cuatro mil ciento sesenta y cuatro diezmillonésimas) tienda a cero.

4. No sería justo terminar estas notas sin subrayar de una vez que el método propuesto por Pelévina no está concebido para señalar cuáles han de ser las condiciones probabilísticas en que deben hallarse dos o más palabras emparentadas etimológicamente y cuáles han de ser las que produzcan el efecto contrario, sino al revés: dadas unas condiciones probabilísticas, extrae unos indicios de vinculación etimológica. La correlación entre probabilidad y parentesco etimológico no existe como una implicación bilateral entre ambos factores, sino como implicación unilateral. Y es que no cabe en modo alguno negar la posibilidad de una etimología común entre estructuras fonológicas muy frecuentes y, por tanto, muy probables. El hecho de que existan no constituye, pues, ninguna limitación al método de Pelévina ni, en consecuencia, un reparo por nuestra parte.

Además, nuestro análisis se ha orientado más bien a aspectos prácticos y operativos aplicados al castellano. Esto podría levantar alguna sospecha sobre la justedad de esta transferencia a partir del ruso, pero creemos que no vale la pena examinar el asunto ni siquiera para defender el carácter teórico y, por tanto, general de la propuesta de Pelévina. En ninguna lengua cabe imaginar que las leyes evolutivas modifiquen o hayan modificado los étimos anárquicamente, hasta el punto de crear resultados fonológicamente asistemá-

ticos (de ello depende, por cierto, el reconocimiento mismo de las «leyes» evolutivas). La tendencia al equilibrio estructural de cualquier sistema lingüístico garantiza la estabilidad básica que va implícita en aquella propuesta. En este sentido, Pelévina emitió un universal lingüístico.

Pero la conclusión no sólo debe referirse al hecho, en principio contingente, de que las lenguas naturales apenas deparan ilustraciones válidas a la determinación probabilística del parentesco etimológico, sino también al aspecto paradójico (y característico en esta suerte de teoremas lingüísticos) de su propia formulación teórica: si a mayor peculiaridad o rareza fonológica hay mayor certeza de identidad etimológica entre un grupo dado de palabras alternativamente sinónimas, entonces es que el método mismo ofrece tanto más poder de verificación cuanto más pequeña sea su capacidad de aplicación a casos reales.

Y hay que conceder albricias a la feliz casualidad de que Pelévina no tropezara con un contraejemplo, sino con un ejemplo. De otro modo, acaso le hubiera faltado la convicción necesaria para formular su interesante principio.

Universidad de Barcelona.

RAMÓN CERDÀ