Capítulo 5
TEORÍA DE LA INFORMACIÓN Y MEDIDAS DEL CAMBIO LINGÜÍSTICO

Todo lo medible se debe medir; lo que no es medible debe hacerse medible.
GALILEO GALILEI

En realidad, las leyes del azar son con frecuencia un modelo mejor de la ignorancia que las leyes de la lógica lo son del pensamiento organizado.
B. MANDELBROT

5.1. Introducción a la teoría de la información

Uno de los desarrollos teóricos más prolíficos del siglo XX ha sido la teoría de la información. Iniciada por el ingeniero eléctrico Claude E. Shannon en su histórico artículo[1] de 1948, la teoría ha encontrado amplia aplicación en muchos y muy diversos campos: en las telecomunicaciones, en la programación, en la física pura y marginalmente en lingüística y psicología. El objetivo de este capítulo es aplicarla a la lingüística de una manera diferente a como se había hecho hasta ahora.

A parte de las aplicaciones lingüísticas, que más adelante discutiremos, hay otras disciplinas colaterales a la misma dentro de las cuáles la teoría de la información ha hecho sus aportaciones como en **criptología** y en **psicología**. Dentro de la criptología (aplicable al desciframiento de textos antiguos) la teoría de la información ha demostrado que la cantidad de información por signo está íntimamente ligada a la posibilidad de descifrar criptogramas: cuanto menor es esta cantidad más sencillo es el desciframiento y se necesita menor cantidad de material para el desciframiento. Así pues dentro de ciertos límites, la teoría se vuelve predictiva y cuantitativa, proporcionando los medios de calcular que cantidad de textos deben interceptarse en cierta lengua con el objeto de asegurar la existencia y la unicidad de una solución al criptograma o texto cifrado. También los psicólogos han encontrado relaciones interesantes entre la cantidad de información (cuantificada según la fórmula de Shannon) contenida en un estímulo y el tiempo de reacción al estímulo. Por ejemplo, en un experimento[2] se colocan cuatro luces y cuatro pulsadores asociados; las luces se encienden y apagan al azar con probabilidades p_1, p_2, p_3, p_4 y se pide a un individuo que apriete los botones correspondientes después de que una luz se apague, tan rápidamente como sea posible. El resultado de este experimento es que el tiempo medio de reacción requerido t_{reac} se incrementa linealmente con la cantidad de información reportada por las luces, es decir:

$$t_{reac} = t_0 + a \cdot I_{Shannon}$$

(siendo t_0 y a números constantes; $I_{Shannon}$: información computada por la fórmula de Shannon, es decir, $I_{shannon} = -(p_1 \ln p_1 + p_2 \ln p_2 + p_3 \ln p_3 + p_4 \ln p_4)$). Este resultado sugiere una connexión intrínseca entre al manera en que los seres humanos procesan la información y la fórmula teórica de Shannon; hecho crucial para la **psicolingüística** de corte matemático.

Cuantificación del cambio lingüístico mediante la teoría de la información

En su artículo original Shannon dedujo una fórmula matemática que da una medida de la **cantidad de información** o **imprevisibilidad** asociada a un proceso de elección entre posibilidades con diferentes probabilidades de ocurrencia[3]. De esta manera podemos evaluar, por

---

[1] Shanon, C. E. (1948): "The Mathematical Theory of Communication", Univetisty of Illinois Press, pp. 3-28.
[2] Holzmüller, W, (1984): *Information in Biological Systems*.
[3] Algunos autores, como Mackay, prefieren llamar **imprevisibilidad** a la magnitud medida por la fórmula de Shanon; está muy claro que la imprevisibilidad de una situación es igual a la cantidad de información necesaria para determinar por completo la elección de una de las posibilidades.

ejemplo, la cantidad de información por fonema de un fragmento leído a partir de las diversas probabilidades de ocurrencia de los diferentes fonemas. Armados con esta medida de la cantidad de información podemos abordar muy diversos problemas. Dentro de la lingüística la teoría de la información sólo se había aplicado a dos problemas: el estudio de la **redundanciaen el lenguaje**[4] (Shannon, 1951) y la **ley de distribución de frecuencias de las palabras**[5][6] (Zipf, 1949; Mandelbrot, 1961).

En este capítulo se pretende aplicar la fórmula de Shannon a la lingüística histórica de una manera que proporcione una medida efectiva del cambio lingüístico. Ilustremos la idea básica con un caso concreto: supongamos, por ejemplo, que tomamos dos lenguas románicas como el francés y el rumano, y pretendemos **cuantificar su divergencia** o disimilitud con respecto al latín. Esto se conseguiría midiendo la "latinicidad" (en los diferentes niveles: fonético-fonológico, morfosintáctico y léxico-semántico) del rumano y del francés. El hecho importante aquí es que la teoría de la información nos permite construir una medida razonable de la "latinicidad" de una lengua. Las ideas claves para lograr esta medida son dos. Por una parte, una persona que conozca el francés o el rumano puede aprovechar este conocimiento para aprender el latín más fácilmente; esto se debe a que hay ciertas correlaciones entre la estructura de estas lenguas y la del latín. Por otra parte, una persona que únicamente hablase latín no podría comprender el francés o el rumano, sin cierto entrenamiento previo; esto se debe a que de alguna manera se han añadido elementos nuevos que no son predictibles a partir únicamente del latín, es decir, se ha añadido información nueva a la base latina original. Ponderando convenientemente la información correlacionada con la información nueva esbozaremos una medida de la "latinicidad" de la siguiente manera: tomemos una muestra significativamente grande de palabras latinas y de sus equivalentes en otra lengua románica; calculemos a continuación, por ejemplo, mediante la fórmula de Shannon las cantidades de información por fonema I_R (para la lengua románica) y I_L (para el latín) además de la información no predictible a partir del latín de la lengua románica[7] que designaremos por $I_{R|L}$. Es evidente que se cumplirá que (más adelante daremos una demostración) $I_{R|L} \leq I_R$, siendo tanto $I_{R|L}$ tanto más alto cuanto menos próxima sea la lengua al latín. Así pues una buena medida de la "latinicidad" a nivel fonético-fonológico C_{fon} será:

$$C_{\text{fon}}\text{: "latinicidad" fonológica} = 1 - \frac{I_{R/L}}{I_R} \qquad (1 \geq C_{\text{fon}} \geq 0)$$

cuanto mayor es el C_{fon} entre una lengua románica y el latín, tanto mayor es su similitud a nivel fonológico con el latín[8]. El procedimiento es general y, de hecho, podemos evaluar la "latinicidad" de cualquier lengua: si evaluamos la "latinicidad" de una lengua como el japonés es previsible que C_{fon} esté próxima a cero, mientras que el inglés tendría un valor moderado debido al elevado número de palabras prestadas del latín que contiene. Obsérvese que por la propia definición de C_{fon} siempre encontramos que su valor está comprendido entre 0 y 1.

El procedimiento puede generalizar a dos lenguas cualesquiera A y B, relacionadas genéticamente o no, aunque el segundo caso aparecen ciertos problemas técnicos que comentaremos más adelante. También se puede generalizar el procedimiento a los demás niveles de la lengua (morfosintáctico y léxico-semántico) aunque para esto debemos recurrir a la lingüística estructural y a la caracterización mediante rasgos distinctivos. Para dar una medida adecuada del cambio lingüístico se requerirán por tanto una serie de índices: C_{fon}, C_{mor}, $C_{\text{lex-sem}}$, (y otros que mencionaremos más adelante), cada uno de los cuales medirá aspectos parciales del cambio.

---

[4]Shanon, E. C. (1949): "Prediction and Entropy of Printed English", *Bell System Tech. J. 30,* pp 50-64.
[5]Zipf, G. K. (1949): *Human Behaviour and the Principle of Least Effort*, Addison-Wesley, Cambride, Massachusetts.
[6]Mandelbrot, B. (1961): "On the theory of word frecuencies and on related markovian models of discurse", en *Structure of language and its mathematical aspects* (volumen dirigido por R. Jakobson), Providence, *American Mathematical Society*.
[7]Más adelante se darán las fórmulas y procedimientos explícitos para estos cálculos. También discutiremos detalladamente los problemas técnicos y conceptuales que se presentan, como por ejemplo, la no-equivalencia exacta de formas léxicas, cómo establecer las correspondencias entre lenguas no emparentadas, etc.
[8]La medida que acabamos de presentar es una medida de la información fonológica preservada (o una medida de la correlación) entre dos variedades diacrónicamente relacionadas (en este caso del latín y una lengua románica); que por la propia definición siempre estará comprendida entre 0 y 1 (o si se quiere entre el 0% y el 100%).

Connexiones con la glotocronología

La medida del cambio aquí desarrollada constituye una innovación en el uso de ciertos métodos matemáticos de peso aplicados a la lingüística. Sin embargo, el uso de cálculos para medir el cambio lingüístico no es una idea nueva; la **glotocronología** es una teoría desarrollada por Swadesh y Lee, que trata de cuantificar el tiempo de separación de dos lenguas relacionadas genéticamente (lenguas "hijas") de su antecesor común (lengua "madre" o protolengua). La teoría de Swadesh también llamada lexicoestadística, no aceptada universalmente y muy controvertida, resulta excesivamente simplista ya únicamente analiza el nivel léxico y utiliza hipótesis difícilmente comprobables y que verosímilmente podrían descartarse. No obstante, la glotocronología conduce muchas veces a resultados razonables cuando se compara con los hechos constados.

Por ejemplo, los cálculos glotocronológicos basados en el porcentaje de coincidencias en el vocabulario de las lenguas románicas y el latín predicen que la separación entre ellas debió empezar hace unos 1500 o 1700 años, fecha que coincide con la última época del imperio cuando suponemos empezó a hacerse más evidente la diferenciación dialectal del latín. Otros cálculos glotocronológicos sugieren que el inglés y el holandés empezaron a divergir en el siglo IX, el alemán del inglés en siglo VI, fechas que también concuerdan aproximadamente con las migraciones germánicas hacia Inglaterra.

La teoría desarrollada en este capítulo es una superación de los métodos estadísticos usados en glotocronología, por varias razones. En primer lugar permite analizar el cambio en sus diferentes niveles (y no se basa únicamente en el nivel léxico). En segundo lugar puede predecir si un determinado cambio potencial se vería favorecido o impedido[9]. Y en tercer lugar utiliza magnitudes intrínsecas con un significado teórico claro[10], y no simplemente extrapolaciones basadas en medias estadísticas como en el caso de la glotocronología.

Connexiones con la entropía física

Otra de las virtudes de la medida aquí desarrollada es que tiene importantes analogías formales con una magnitud muy importante en física llamada **entropía**. El universo parece comportarse de tal manera que cualquier cambio espontáneo aumenta la entropía total del mismo; así observamos ciertos procesos como que el humo de un cigarrillo se dispersa espontáneamente en el aire de una habitación, o que dos cuerpos en contacto y con una diferencia de temperatura tienden a disminuirla (es decir, pasa calor del cuerpo caliente al más frío). Estos procesos comportan un aumento de la entropía. Los procesos contrarios, que son los que veríamos en una película pasada de atrás a adelante, comportan una disminución de la entropía del universo y nunca los hemos observado a pesar de no contradecir ningún principio básico de la física. Por esta razón se pensó en introducir un nuevo principio básico conocido como "segundo principio de la termodinámica" que prohibe los procesos en los que la entropía total del universo disminuye. Algunos físicos intuyeron que este principio iba más allá de la física y lo aplicaron con éxito a otros campos como la biología y la economía. En lingüística, por su parte, encontramos que ciertos hechos espontáneos y muy frecuentes en las lenguas del mundo como la **asimilación fonética** y la **analogía**, comportan un aumento del análogo lingüístico de la entropía (aunque no sea cierto en general que las lenguas evolucionen en el sentido de aumentar su entropía).

5.2. La fórmula de Shannon. Medida de la información introducida

Supongamos ahora una situación futura incierta en la que sólo sabemos pueden suceder P_0 cosas diferentes, en general, con probabilidades diferentes. Supongamos ahora que deseamos hacer una elección o tomar una decisión estratégica sobre que hacer en esta situación. En estas condiciones la

[9]Desde el punto de vista de la optimización de los recursos de la lengua; criterio éste que suele coincidir con la dirección real de los cambios (asimilación fonética, analogía morfológica, etc ...).

[10]La principal magnitud que trataremos es la imprevisibilidad de Shanon cuya definición coincide con la de entropía física cuyo significado ha sido puesto en claro extensivamente por la física teórica desde más de un siglo.

función de Shannon nos da una medida de la **imprevisibilidad**[11] de la situación, o vista de otra manera de la **cantidad de información** adicional necesaria para hacer la situación completamente previsible.

Imaginemos ahora un caso relacionado con la lingüística. Supongamos que se nos pide que adivinemos una palabra castellana incógnita de cuatro letras y, para facilitarnos las cosas, nos informan de cuál es la primera letra. Esta información adicional puede tener un valor variable, si la primera letra es una Z se nos da una información I_1 mucha más valiosa, por ejemplo, de la información I_2 consistente en decirnos que la primera letra es M; ya que en el primer caso nos será mucho más fácil dar con la palabra, mientras que en el segundo el número de posibilidades es bastante mayor. En este caso la fórmula de Shannon, que deduciremos a continuación, también puede cuantificar cuanto más valiosa es I_1 que I_2 (téngase en cuenta que I_1 reduce la imprevisibilidad en mayor grado que I_2).

Deducción de la fórmula de Shannon[12]

Antes de escribir la fórmula general de Shannon conviene analizar un caso particular muy ilustrativo, que dejará más clara la interpretación de la fórmula general. Consideremos una situación compuesta por la coincidencia simultánea de dos situaciones simples independientes: pongamos que en la primera situación simple pueden darse Ω_1 casos posibles (todos ellos equiprobables) y en la segunda Ω_2. Como cada uno de los Ω_1 casos de la primera situación puede darse en conjunción con cada uno de los Ω_2 de la segunda, el número de casos en la situación compuesta será el producto $\Omega_1 \cdot \Omega_2$, la imprevisivilidad en la situación compuesta H_{comp} deberá ser igual a la suma H_1+H_2 (por ser independientes ambas situaciones). En estas condiciones se demuestra[13] que la única relación posible para que ambas cosas se cumplan simultáneamente es que se cumplan las ecuaciones:

$$[1] \quad H_1 = k \ln \Omega_1 \qquad\qquad H_2 = k \ln \Omega_2$$

(siendo k una constante[14] que depende de las unidades escogidas para la cantidad de información y ln el logaritmo natural); ya que de esta manera:

$$[2] \quad H_{comp} = k \ln \Omega_{comp} = k \ln (\Omega_1 \cdot \Omega_2) = k (\ln \Omega_1 + \ln \Omega_2) = H_1 + H_2$$

Esta última propiedad, se conoce como **aditividad de la información**. Particularicemos ahora un poco más este ejemplo, un tanto abstracto; pongamos que tratamos de medir la imprevisibilidad asociada a la ocurrencia de número aleatorio de dos cifras. Esta situación puede descomponerse en dos situaciones más simples e independientes: la elección de la primera cifra y la elección de la segunda cifra. La situación compuesta presenta 100 casos equiprobables (00 a 99), mientras que las situaciones tan sólo 10 cada una (0 a 9). Sin embargo, la imprevisibilidad en la situación compuesta es, obviamente, tan solo el doble que en las simples[15], a pesar de que el número de casos es mucho mayor. Con respecto a la unidad de medida de la información utilizaremos el bit que es la unidad resultante cuando se escoge $k = 1 / \ln 2$.

Si ahora introducimos probabilidades podemos escribir la relación entre la imprevisibilidad y el número de casos de una forma fácilmente generalizable. En nuestro caso si tenemos Ω posibilidades equiprobables entonces la probabilidad de cada uno de los casos será $p_i = 1/\Omega$ (i = 1, 2, ..., Ω), por lo que escribiremos[16]:

---

[11] La mayoría de los autores se refieren a la magnitud medida por la fórmula de Shannon como *cantidad de información*. He preferido utilizar la mayoría de las veces el término *imprevisibilidad* ya que el primero continúa sufriendo multitud de implicaciones emocionales debidas a su nombre.

[12] Esta sección presenta una deducción informal de la fórmula de Shannon y puede ser omitida si resulta difícil.

[13] Véase por ejemplo Roman, S. (1992), Brioullin, L. (1962) e incluso textos elementales de cálculo.

[14] Por ejemplo si se escoge $k = 1/\ln 2$ las unidades resultantes son bits y se podrá escribir simplemente como $I = \log_2 W$ ya que $\log_2 x = \ln x / \ln 2$. Si escoge k=1/ln 8 las unidades serán bytes, y si se escoge k igual a la constante de Boltzmann las unidades coincidirán con las de la entropía física.

[15] $I_{comp} = k \ln 100 = k \ln (10^2) = 2k \ln 10 = 2\, I_{simp}$ (recuérdese que $\log_b a^c = c \log_b a$).

[16] Recuérde se la conocida propiedad de los logaritmos: $\log_b (1/x) = - \log_b x$. Para la correcta interpretación de los signos que aparecen en esta ecuación, véase un poco más adelante.

$$H = k \ln \Omega = - k \ln (1/\Omega) = - k \ln p_i = -k \sum_{i=1}^{i=\Omega} p_i \ln p_i$$

fórmula, ésta última, que coincide con el caso general de que $p_1, ..., p_\Omega$ sean diferentes entre sí. De hecho puede demostrarse que toda medida razonable de la cantidad de información es necesariamente de la forma:

[3] $$H(p_1, ..., p_n) = -k (p_1 \ln p_1 + ... + p_n \ln p_n) = -k \sum_{i=1}^{i=n} p_i \ln p_i$$

Es decir, la **imprevisibilidad** es la **media del logaritmo de la improbabilidad** (improbabilidad = 1 / probabilidad). La función de Shannon (también llamada entropía por la mayoría de autores) siempre es mayor o igual que cero. Puede alcanzar su valor mínimo, el cero, cuando un mensaje o situación es completamente determinista (*i.e.* tiene probabilidad 1 o del 100%, y el resto de mensajes probabilidad 0). Intuitivamente esto significa que en una situación determinista a priori no hay ninguna imprevisibilidad y no puede extraerse nueva información; de la misma manera una tautología[17] no aporta nada a nuestro conocimiento del mundo.

Un juego de salón llamado *Veinte preguntas* puede ilustrar muchas de las ideas expuestas. En este juego, una persona piensa en un objeto o persona, mientras los otros jugadores intentan determinar de qué se trata haciendo no más de 20 preguntas cuyas respuestas son "sí" o "no". De acuerdo con la teoría de la información cada pregunta reporta una información menor o igual a $\log_2(2)$, es decir, un bit. La estrategia óptima en este juego, es decir, la que optimiza la información obtenida mediante un número fijo de preguntas, consiste en que los jugadores hagan preguntas tales que dividan al conjunto de posibilidades en dos grupos tan equiprobles como sea posible. Por ejemplo, si se ha establecido por las preguntas anteriores que se trata de adivinar el nombre de un político sería una tontería preguntars a continuación si es hombre o mujer ya que el número de políticos hombres es abrumadoramente mayor. En este caso sería mucho mejor si este político es anterior a una determinada fecha, que divida a los políticos conocidos en dos grupos más o menos iguales. Si los jugadores siguen la estrategia óptima serían capaces de identificar un objeto o persona entre 2^{20} (aproximadamente 1 050 000) posibilidades, lo que corresponde a 20 bits que es la información máxima extraible en veinte preguntas de acuerdo con la teoría de la información.

Por último cabe hacer algunos comentarios sobre la unicidad de medida propuesta. Ya hemos dicho que toda medida razonable lleva a una fórmula como la de Shannon. Entendiéndose aquí por medida razonable una que satisfaga tres siguientes **condiciones razonables** que formuladas técnicamente son[18]:

(i) *$H(p_1, ..., p_n)$ es una función continua de las probabilidades.* Esta condición implica de distribuciones de probabilidad similares dan cantidades de información similares.

(ii) *$H(1/n, ...,1/n) < H(1/(n+1), ...,1/(n+1))$ para todo $n>0$.* Esta condición implica de entre dos situaciones en que todos los casos son igualmente probables, presenta una mayor cantidad de información la que presenta un mayor número de casos.

(iii) *Para un conjunto de valores $b_i>0$ $(i=1,...,k)$, tales que $b_1 + ... + b_k = n$. Se cumple que:*

$$H(\frac{b_1}{n},\ldots,\frac{b_k}{n}) = H(\frac{1}{n},\ldots,\frac{1}{n}) - \sum_{i=1}^{i=k}(\frac{b_i}{n} H(\frac{1}{b_i},\ldots,\frac{1}{b_i}))$$

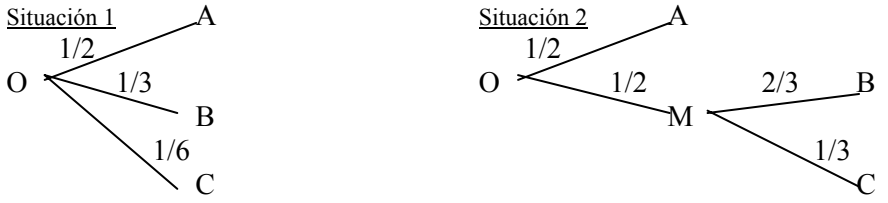
Esta condición técnica proporciona una fórmula práctica para evaluar la cantidad de información en el caso de que no haya equiprobabilidad entre los casos[19], sin más que hacer $p_i=b_i/n$. Por otra parte la razonabilidad de la fórmula puede ilustrarse mediante el siguiente esquema:

[17]Es decir, una proposición lógica que es cierta con independencia de la veracidad de las variables como: *p o no-p*, la probabilidad de certeza para una proposición así es 1. E igualmente para una contradicción como: *p y no-p*, cuya probabilidad de certeza es 0.

[18]Roman, S. (1992): *Coding and information theory*, Springer Verlag, New York.

[19]En caso de que todas las p_i sean números racionales, en caso contrario debemos utilizar también la condición (i).

Tabla 5.1: Dos situaciones futuras equivalentes[20].



Desde el punto de vista del resultado final ambas situaciones conducen a los mismos resultados A, B, y C con las mismas probabilidades (números sobre las lineas). Planteadas desde la situación O son equivalentes.

Como las dos situaciones presentadas en el anterior cuadro son equivalentes vistas desde la perspectiva de O, las imprevisibilidades dados por la fórmula de Shannon deberían ser las mismas. Nótese que la *situación 1* es simple, mientras que la *situación 2* es compuesta (una primera situación conduce a A o M; y si se da el segundo caso aparece otra situación que lleva finalmente a B o C). Pues bien la condición (iii) asegura que las imprevisibilidades calculadas serán las mismas (para comprobarlo basta tomar: b_1=3, b_2=2, b_3=1 y n=6).

Cuantificando la información a partir de la imprevisibilidad

Con nuestra medida de la imprevisibilidad estamos en disposición de cuantificar la información aportada por una fuente. Supongamos una situación con una imprevisibilidad asociada de $H_{inicial}$ (por ejemplo la elección de una palabra de n letras) y que se nos proporciona una información I (por ejemplo se nos informa sobre cuál es la primera letra), de tal manera que la imprevisibilidad final después de tener en cuenta esta información resulta ser H_{final}, entonces se cuantifica I como:

[4] $$I = H_{inicial} - H_{final} = -(H_{final} - H_{inicial})$$

en principio I>0 significa que esta información resulta útil ya que disminuye la imprevisibilidad.

5.3. Glotocronología: méritos y defectos.

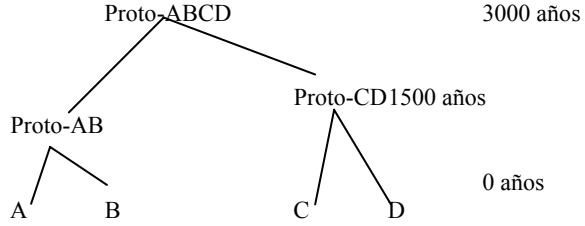
En esta sección vamos esbozar brevemente el método glotocronológico, y vamos a plantear las principales objeciones. Los métodos de la teoría de la información aquí expuestos pueden resolver gran parte de los problemas que presenta la glotocronología; sin embargo, intentaremos ir un poco más allá contruyendo un nuevo método. Por su parte, la glotocoronología es un desarrollo de la lingüística histórica relativamente reciente cuyo objeto es la medida de la divergencia de dos lenguas emparentadas genéticamente. La idea básica es muy simple. Se empieza con la observación general de que cuanto mayor es el intervalo temporal que separa a los miembros de la familia de su antecesor común (lengua "madre"), tanto mayor es la diferencia entre las lenguas "hijas". Por lo que a vocabulario se refiere, las lenguas emparentadas difieren porque algunas palabras dejan de usarse y son reemplazadas por otras, es decir, mediante **cambio léxico**. Así pues dos lenguas que en un tiempo tuvieron vocabularios prácticamente idénticos, cuando podían considerarse dialectos de una misma lengua, difieren cada vez más con el paso del tiempo[21]. La idea se refleja en el siguiente gráfico:

[20]Para calcular la probabilidad de obtenr A, B o C a partir de O, basta con seguir el camino de uno a otro y multiplicar los números que aparecen sobre las lineas.

[21]Esta hipótesis es potencialmente falsa y tan sólo en caso de que el contacto entre ellas se haga cada vez más débil, cosa que no siempre es así: El inglés se ha visto muy influido por el francés y el latín como lenguas de cultura, lo que les ha acercado más a las lenguas románicas.

Tabla 5.2: *El uso de datos léxico-estadísticos para la inferencia lingüística.*

	Coincidencia en el léxico:	tiempo de separación:
A-B	70%	1184 años
A-C	40%	3043 años
A-D	42%	2881 años
B-C	41%	2961 años
B-D	40%	3043 años
C-D	60%	1696 años



Los pioneros de la glotocronología, particularmente Morris Swadesh, estimaron que al menos para un núcleo esencial del vocabulario la tasa de substitución de vocabulario podía considerarse constante[22]; esta hipótesis funciona razonablemente bien para periodos de tiempo largos. Swadesh escogió un conjunto básico de prueba una serie de términos del vocabulario nuclear que no estuvieran marcados culturalmente[23]. El procedimiento seguido por los glotocornologistas, cuando tratan de determinar la separación con el tiempo de dos lenguas relacionadas, es escribir la lista de términos equivalentes (dada en la siguiente tabla) de las lenguas consideradas:

1 yo	18 persona	35 rabo	52 corazón	69 estar de pie	86 montaña
2 tú	19 pez	36 pluma	53 hígado	70 dar	87 rojo
3 nosotros	20 pájaro	37 pelo	54 beber	71 decir	88 verde
4 este	21 perro	38 cabeza	55 comer	72 sol	89 amarillo
5 aquel	22 piojo	39 oreja	56 morder	73 luna	90 blanco
6 quién	23 árbol	40 ojo	57 ver	74 estrella	91 negro
7 qué	24 semilla	41 nariz	58 oír	75 agua	92 noche
8 no	25 hoja	42 boca	59 saber	76 lluvia	93 caliente
9 todo	26 raíz	43 diente	60 dormir	77 piedra	94 frío
10 muchos	27corteza	44 lengua	61 morir	78 arena	95 lleno
11 uno	28 piel	45 garra	62 matar	79 tierra	96 nuevo
12 dos	29 carne	46 pie	63 nadar	80 nube	97 bueno
13 grande	30 sangre	47 rodilla	64 volar	81 humo	98 redondo
14 largo	31 hueso	48 mano	65 caminar	82 fuego	99 seco
15 pequeño	32 grasa	49 vientre	66 venir	83 ceniza	100 nombre
16 mujer	33 huevo	50 cuello	67 mentir	84 quemar	
17 hombre	34 cuerno	51 senos	68 sentarse	85 camino	

y anotar los pares que pueden considerarse coincidentes por su similaridad (teniendo en cuenta las correspondencias fonéticas regulares). Estos pares coincidentes se presuponen se remontan al antecesor común, mientras que las palabras que tienen diferente forma en las dos lenguas diferirían a causa de que la palabra original se habría perdido en una lengua o en ambas lenguas. El número de pares coincidentes es ya una medida de la proximidad lingüística, sin embargo, la glotocronología va algo más allá tratando de determinar el tiempo se separación. En el estudio original[24], varios pares de lenguas se compararon, entre ellas el antiguo anglosajón y el inglés medio, el latín de Plautus y el francés antiguo, el chino clásico y el moderno mandarín o *guanhua*, entre otros. A partir de los datos obtenidos se concluyó que el porcentaje de retención era del 86% por milenio. Con este valor como guía puede estimarse el tiempo de separación entre dos lenguas emparentadas como:

[22]Ya comentamos en la introducción del capítulo lo controvertido de esta hipótesis. Cuando factores políticos o históricos en una comunidad aparece el bilingüismo los cambios léxicos empiezan a ser mucho más frecuentes.
[23]Como es sabido los elementos del vocabulario periférico son propios de cada cultura y varían considerablemente con el tiempo. Aun así la elección de Swadesh es muy discutible ya que incluye palabras como 'verde', 'amarillo', 'árbol' o 'nadar' que no existen en todas las lenguas del mundo (por lo que están marcadas culturalmente).
[24]Lees, R.B. (1953): "The basis of glottochronology", *Language* 29, p.113-125.

$$\textit{tiempo de separación} = 1000\ (\ \ln p\ /\ \ln p_0\ )$$

Siendo p el porcentaje de coincidencias expresado en tanto por 1, y p_0 una constante que vale p_0=0,86 cuando comparamos una lengua con otra que deriva de ella, y $p_0 = 0{,}74 = 0{,}86^2$ cuando comparamos dos lenguas que derivan de una antecesor común[25].

En mi opinión el método glotocronológico presenta varios problemas. En primer lugar, tan sólo se considera el cambio léxico y sin tener en cuenta la disimilitud fonética entre dos términos que designan el mismo concepto. Por ejemplo, si consideremos las formas léxicas para 'corazón' en diversas lenguas indoeuropeas, vemos profundas diferencias entre algunas lenguas a pesar de la coincidencia léxica (establecida por el método comparativo).

Tabla 5.3: *La evolución de *krd 'corazón' del indoeuropeo hasta las lenguas históricas*.

	II a.C.	II a.C.	I a.C.	I d.C.		
(indoeuropeo)	**kr̥d / *kred*	*>*kred*	*>*kʲrad*	*>śrad-*	[ʃrəd]	(sánscrito)
(indoeuropeo)	**kr̥d*	*>*k^{a}rd*	*>*kard*	*>kard-*	[kard]	(griego)
(indoeuropeo)	**kr̥d*	*>*k^{o}rd*	*>*kord*	*>cord-*	[kord]	(latín)
(indoeuropeo)	**kr̥d / *kerd*	*>*k^{h}erd*	*>*hɛrt*	*>heart*	[hɑ:t]	(inglés)

En indoeuropeo las formas **kerd*, **kr̥d*, **kred* se utilizaban según el caso gramatical.
r̥ = r silábica, es decir, funcionando como vocal o centro de sílaba.

Todas ellas a pesar de sus diferencias fonéticas (*cf*. sánscrito [ʃrəd-], inglés [hɑ:t], latín [kord-]) deben considerarse coincidentes, porque derivan del mismo vocablo indoeuropeo. Esta manera de contar es demasiado taxativa (blanco/negro, sí/no, coincide/no-coincide); no basta decir que dos formas léxicas "coinciden" si no que debe precisarse de alguna manera "cuanto" coinciden. Una mejora del método sería asignar un número (entre 0 y 1) a cada par de términos (0 = ningún tipo de coincidencia, 1 = coincidencia perfecta) y sumar el total, en lugar de sumar el número de coincidencias léxicas (esta segunda posibilidad equivale a hacer únicamente asignaciones 0 o 1 y sumar). Las ecuaciones desarrolladas en el presente capítulo establecerán un método objetivo para asignar el grado de coincidencia. Esta mejora permite tener en cuenta el grado de similitud fonética y no únicamente el número de "coincidencias" léxicas. Ya hemos mencionado otras críticas menores a la glotocronología, pero en mi opinión la no coincidencia exacta de formas léxicas era la más problemática, y puede ser corregida en la forma indicada.

El objetivo de las siguientes secciones es construir métodos independientes a la glotocronología para medir el cambio lingüístico, sin embargo, algunas de las soluciones podrían ser aplicadas para corregir las deficiencias del método glotocronológico.

5.4. Primera aproximación a la medida del cambio lingüístico

En la lengua hablada las unidades elementales son los fonemas, mientras que el lenguaje escrito consiste en secuencias de letras. Así las ecuaciones y conceptos introducidos en la sección anterior estamos ya en disposición de evaluar la cantidad de información por unidad elemental tanto de la lengua escrita como hablada.

Para ilustrar esto consideraremos primero el problema de evaluar la cantidad de información en una oración esctita. Este problema, intrincado y de gran importancia práctica, ha sido discutido cuidadosamente por C.E. Shannon; sin embargo, a pesar de los multiples esfuerzos por parte de diversos autores aún estamos lejos de haber obtenido una solución completa y rigurosa (a causa del problema de la redundancia[26] en el lenguaje sobre el que volveremos más adelante). Nuestras unidades elementales, en el caso de la lengua escrita, son las letras; si tomamos como referencia el

[25]En este último caso suponemos que ambas lenguas divergen a partir del antecesor común de manera independiente, en caso de que haya un contancto apreciable entre ellas $0{,}74 < p_0 < 0{,}86$.
[26]Los problemas son sobre todo de tipo práctico; se acusa especialmente la falta de datos estadísticos suficiente para cuantificar la redundacia. Sin embargo, los problemas teóricos no parecen excesivamente graves.

inglés escrito (que es la lengua escrita para la existe un mayor número de datos estadísticos) el alfabeto constaría de 27 símbolos (26 letras + el espacio "en blanco", también llamado fonema nulo). Si todos estos símbolos fueran igualmente probables la ocurrencia de uno de ellos en un texto aportaría una información de $\log_2$ 27 bits[27], y por tanto una oración compuesta por G símbolos (letras y espacios) comportaría una información de:

$$I = G \log_2 27 \text{ (bits)} \qquad \text{o bien} \qquad i = \log_2 27 = 4{,}76 \text{ (bits/símbolo)}$$

Sin embargo, esta solución no es satisfactoria ya que sabemos que ciertas letras son más frecuentes que otras, y por tanto, la imprevisibilidad no era tan grande. La imprevisibilidad eliminada por la ocurrencia de cierto símbolo en general es menor de 4.76 bits. Si tenemos en cuenta las probabilidades[28] de aparición de las diferentes letras en un texto inglés, podemos mejorar al anterior medida:

$$I = -G \sum_{j=1}^{j=27} p_j \log_2 p_j \quad \text{o bien} \quad i = -G \sum_{j=1}^{j=27} p_j \log_2 p_j = 4{,}03 \text{ (bits/simbolo)}$$

Dónde la p_j denota la probabilidad de j-ésimo símbolo (los datos han sido tomados de Brioullin (1962)). Esta medida sería ya correcta si la ocurrencia de los símbolos fuera independiente, sin embargo, en las lenguas naturales las cosas no funcionan así en general la ocurrencia de un fonema modifica las probabilidades a priori para el segundo fonema. Por ejemplo, en japonés es mucho más probable que a una consonante le siga una vocal que no otra consonante, ya que los grupos de varias consonantes son extremadamente raros. El problema puede resolverse cuando estudiemos las correcciones por redundancia, sin embargo, la medida que acabamos de obtener para la información por símbolo ya es suficientemente buena para los cálculos. Vamos a introducir ahora una medida de la **correlación fonológica** entre dos lenguas. Dadas dos palabras $W^{(A)}$ y $W^{(B)}$ de dos lenguas, emparentadas genéticamente, estaremos en disposición de asignar un índice entre 0 y 1del grado de correlación.

Probabilidad e información condicional

Previamente necesitamos introducir los conceptos de probabilidad conjunta y probabilidad condicionada. Las introduciremos a través de un sencillo ejemplo; supongamos que queremos ver la correlación en una determinada comunidad entre el sexo de una persona y el tipo de profesión que desempeña (Intuitivamente esperamos encontrar que en ciertas profesiones predominen o bien hombres o bien mujeres, y que análogamente para un determinado sexo sean más comunes unas profesiones que no otras). Denotaremos los sexos por M (= masculino) y F (= femenino) y a las profesiones por P_1, P_2, ..., P_k. A continuación evaluemos las siguientes probabilidades:

$p_{SP}(M,P_i)$	probabilidad de que un individuo al azar sea hombre y tenga la profesión P_i.
$p_{SP}(F,P_i)$	probabilidad de que un individuo al azar sea mujer y tenga la profesión P_i.
$p_S(F)$ / $p_S(M)$	probabilidad de que un individuo al azar sea mujer / hombre (sin importar la profesión).
$p_P(P_i)$	probabilidad de que un individuo tenga la profesión P_i (sin importar el sexo).

A la probabilidad $p_{SP}(,)$ se llama **probabilidad conjunta** ya que examina ambos factores (sexo / profesión) al mismo tiempo, mientras que $p_S( )$ y $p_P( )$ son probabilidades simples (tal y como hemos definido el problema es evidente que se cumplirá que: $p_P(P_i) = p_{SP}(M,P_i) + p_{SP}(F,P_i)$, $p_S(F) = p_{SP}(F, P_1) + ... + p_{SP}(F,P_k)$, etc.).

A partir de las probabilidades ya calculadas estamos ya en condiciones de evaluar al correlación entre sexo y profesión, mediante las siguientes **probabilidades condicionadas**, $p_{P|S}(Pi|M)$ probabilidad de que un individuo tenga la profesión P_i si sabiendo que se trata de un hombre (análogamente $p_{P|S}(P_i|M)$), $p_{S|P}(F|P_i)$: probabilidad de que se trate de una mujer si su profesión es P_i,

[27] Ya que es en esta cantidad precisamente en la que se disminuye la imprevisibilidad de la situación previa.
[28] Puede calcularse tomando un texto en inglés suficientemente largo o puede consultarse la bibliografía, véase por ejemplo, Roman, S. (1992) o Brioullin, L. (1962).

. La teoría elemental de la probabilidad nos dice como calcular la probabilidad condicionada conociendo la probabilidad conjunta y las probabilidades simples:

$$p_{P|S}(\mathrm{P_i}|\mathrm{M}) = \frac{p_{SP}(\mathrm{M},\mathrm{P_i})}{p_{SP}(\mathrm{M},\mathrm{P_1})+\ldots+p_{SP}(\mathrm{M},\mathrm{P_k})} = \frac{p_{SP}(\mathrm{M},\mathrm{P_i})}{p_S(\mathrm{M})}$$

$$p_{S|P}(\mathrm{M}|\ \mathrm{P_i}) = \frac{p_{SP}(\mathrm{M},\mathrm{P_i})}{p_{SP}(\mathrm{M},\mathrm{P_i}) + p_{SP}(\mathrm{F},\mathrm{P_i})} = \frac{p_{SP}(\mathrm{M},\mathrm{P_i})}{p_P(\mathrm{P_i})}$$

Apliquemos estas ecuaciones a los datos de la siguiente tabla, en se representa el reparto (en tanto por 1) de hombres y mujeres entre tres profesiones hipotéticas P_1, P_2, y P_3; se pretende analizar su **correlación** con la variable sexo:

	M	**F**	
P_1	0,36	0,04	*0,40*
P_2	0,1575	0,1925	*0,35*
P_3	0,1825	0,065	*0,25*
	0,70	*0,30*	

Si calculamos por ejemplo $p(\mathrm{M}|\mathrm{P_1})$ = 0,36 / 0,4 =0,9 vemos que en esta profesión predominan los hombres; igualmente el cálculo de $p(\mathrm{M}|\mathrm{P_2})$ = 0,1575 / 0,35 = 0,45 y $p(\mathrm{M}|\mathrm{P_3})$ = 0,1825 / 0,25 = 0,73, nos dice que los hombres son minoría en la primera y son claramente mayoritarios en la segunda. Igualmente las cantidades $p(\mathrm{P_1}|\mathrm{M})$, $p(\mathrm{P_2}|\mathrm{F})$, etc. nos proporcionan información útil.

Ahora podemos escribir la imprevisibilidad asociada a las situaciones simples (probabilidades simples $p_S(\ )$ y $p_P(\ )$) y también la asociada a la situación compuesta (probabilidad conjunta $p_{SP}(\ )$), sin más que sumar para todos los casos posibles:

$H_S = -k(\ p_S(\mathrm{M}) \ln p_S(\mathrm{M}) + p_S(\mathrm{F}) \ln p_S(\mathrm{F}))$ (imprevisibilidad referente al sexo)
$H_P = -k(\ p_P(\mathrm{P_1}) \ln p_S(\mathrm{P_3}) + p_P(\mathrm{P_2}) \ln p_S(\mathrm{P_2}) + p_P(\mathrm{P_3}) \ln p_S(\mathrm{P_3}))$ (imprevisibilidad referente a la profesión)
$H_{SP} = -k(\ p_{SP}(\mathrm{M},\mathrm{P_1}) \ln p_{SP}(\mathrm{M},\mathrm{P_1}) + p_{SP}(\mathrm{M},\mathrm{P_2}) \ln p_{SP}(\mathrm{M},\mathrm{P_2}) + \ldots + p_{SP}(\mathrm{F},\mathrm{P_3}) \ln p_{SP}(\mathrm{F},\mathrm{P_3}))$ (imprev. total)

Vemos que en nuestro caso H_S = 0,8813 bits, H_P = 1,5589 bits y H_{SP} = 2,2981, por lo que $H_S + H_P \neq H_{SP}$. En general puede demostrarse[29] dadas dos variables X e Y con sus respectivas probabilidades simples y conjunta sucede que $H_X + H_Y > H_{XY}$. En lugar de dar una demostración formal de este hecho daremos una interpretación intuitiva. Dado que existe en nuestro ejemplo hay correlación entre sexo y profesión si se nos dice, por ejemplo, es hombre (lo que corresponde a una información $I_1 = H_S$) hemos de pensar que también se nos está aclarando algo sobre su profesión (sabemos por ejemplo que lo más probable es que trabaje en P_1). Si a continuación se nos informa sobre su profesión ($I_2 = H_P$) también se nos aclara algo sobre su sexo (ya que entonces sabremos si es más probable que sea hombre o mujer). Es como si con I_1 e I_2 estuviéramos midiendo una parte de la información dos veces, por lo que el resultado numérico $I_1 + I_2$ (= $H_S + H_P$) es mayor que $I_{min} = H_{SP}$.

Como en general $H_X + H_Y \neq H_{XY}$ (a menos que las variables sean independientes), nos interesaría disponer de una magnitud que sumada H_X (análogamente para H_Y) nos diera precisamente H_{XY}, esta cantidad es lo que se llama **información condicional** $H_{Y|X}$ (analog. $H_{X|Y}$). La información $H_{Y|X}$ es la cantidad de información asociada a la variable Y y que no nos proporciona H_X por sí sola. Así si se nos da H_X y $H_{Y|X}$ cada porción de la información se mide una sóla vez (y no como antes) de tal manera que ahora sí: $H_X + H_{Y|X} = H_{XY}$. Por otra parte es sencillo ver que $H_{Y|X}$ puede calcularse mediante una fórmula a la información ordinaria[30]:

[29] Véase por ejemplo Roman, S. (1992) o Brioullin, L. (1962).
[30] Aunque puede resultar más práctico calular mediante: $H_{Y|X} = H_{XY} - H_X$.

[5] $$H_{Y|X} = -k\sum_{i,j} p_{XY}(\mathrm{i},\mathrm{j}) \ln p_{Y|X}(\mathrm{j}|\mathrm{i})$$

Dónde la suma se extiende a todos los valores posibles de X e Y.
Ahora estamos en condiciones de demostrar que C_{fon} está comprendido entre 0 y 1, puesto que:

$$H_X+H_Y \geq H_{XY} = H_X + H_{Y|X} \geq 0 \;\Rightarrow\; H_Y \geq H_{Y|X} \geq 0 \;\Rightarrow\; 1 \geq H_{Y|X}/H_Y \geq 0$$

a partir de lo cuál es evidente que C_{fon} está comprendido entre 0 y 1.

Supongamos ahora una situación futura incierta Ahora podemos trasladar todo lo anterior, a la correlación entre fonológica entre dos lenguas A y B, emparentadas genéticamente. Calculemos para cada una de las lenguas las probabilidades de aparición de los diferentes fonemas (el sistema fonológico y el número de fonemas no tiene porqué coincidir) es decir $p_A(\mathrm{f_i})$ y $p_B(\mathrm{f_j})$ y para listas de términos análogos y para una misma posición de palabra calculemos la probabilidad conjunta $p_{AB}(\mathrm{f_i},\mathrm{f_j})$ de que aparezca $\mathrm{f_i}$ en la lengua A mientras que en el término análogo de la lengua B y en la misma posición aparezca el fonema $\mathrm{f_j}$. Al igual que en caso anterior se cumplirán las siguientes ecuaciones:

$$p_A(\mathrm{f_i}) = \sum_{j=1}^{j=n} p_{AB}(\mathrm{f_i},\mathrm{f_j}) \qquad p_B(\mathrm{f_j}) = \sum_{i=1}^{i=m} p_{AB}(\mathrm{f_i},\mathrm{f_j})$$

(siendo *m* el número de fonemas en la lengua A, y *n* el número de fonemas en la lengua B). Con estas probabilidades tal y como han sido definidas estamos en disposición de definir la correlación o grado de parentesco entre las lenguas A y B[31], como cociente de información condicionada entre información total. Tan sólo es necesario evaluar las tres magnitudes siguientes:

$$H_A = -k\sum_{i=1}^{i=m} p_A(\mathrm{f_i}) \ln p_A(\mathrm{f_i}) \qquad H_B = -k\sum_{j=1}^{j=n} p_B(\mathrm{f_j}) \ln p_{\mathrm{B}}(\mathrm{f_j})$$

$$H_{AB} = -k\sum_{i=1}^{i=m}\sum_{j=1}^{j=n} p_{AB}(\mathrm{f_i},\mathrm{f_j}) \ln p_{AB}(\mathrm{f_i},\mathrm{f_j})$$

Una vez calculadas estas magnitudes (cuyo cálculo puede ser bastante tedioso) se obtienen fácilmente los valores de $C_{fon}(\mathrm{A}|\mathrm{B})$ y $C_{fon}(\mathrm{B}|\mathrm{A})$ de la siguiente manera:

$$H_{A|B} = H_{AB} - H_B \qquad H_{B|A} = H_{AB} - H_A$$
$$C_{fon}(\mathrm{A}|\mathrm{B}) = 1 - (H_{A|B} / H_A) \qquad C_{fon}(\mathrm{B}|\mathrm{A}) = 1 - (H_{B|A} / H_B)$$

Obsérvese que a menos que $H_A = H_B$, resultará que $C_{fon}(\mathrm{A}|\mathrm{B})$ y $C_{fon}(\mathrm{B}|\mathrm{A})$ serán diferentes. Esta diferencia es lógica si pensamos que conocida la estructura fonológica de la lengua A, la información adicional que nos hace falta para conocer la lengua B es diferente que la información adicional para conocer la estructura de la lengua A a partir de la lengua B. Si $C_{fon}(\mathrm{A}|\mathrm{B})$ y $C_{fon}(\mathrm{B}|\mathrm{A})$ no difieren demasiado una buena medida de la correlación fonética entre las lenguas A y B, es la media aritmética de ambos.

5.5. Cálculos explícitos y primeros resultados numéricos

Con estas ideas en mente podemos empezar a calcular y ver comprobar la utilidad del método expuesto. Mi primer intento fue cuantificar la divergencia de 5 lenguas indoeuropeas con respecto al indoeuropeo clásico[32]. Para ello tomé una lista de 75 vocablos indoeuropeos, cuya

[31] En realidad, tan sólo el grado de parentesco entre las dos listas de términos análogos escogidas. Diferentes listas comportarán diferentes correlaciones, sin embargo, es de esperar que si las listas son representativas y suficientemente extensas el valor de numérico de la correlación será aproximadamente el mismo.
[32] Se entiende aquí el indoeuropeo reciente tal y como se reconstruye mediante el método comparativo clásico.

reconstrucción no ofreciera especiales problemas, y sus equivalentes en sánscrito, griego, latín, germánico[33] y baltoeslavo[34].

Cuando estaba cerca de terminar estos cálculos cayó en mis manos el trabajo del lingüista checo H. Kucera[35] sobre la **isotopía fonemática**. Kucera, al igual que nosotros, había intentado definir una medida de a similitud fonológica de dos elnguas mediante un índice comprendido entre 0 y 1, al que llamó isotopía. Aunque el método de Kucera difiere en lo esencial del nuestro hay similitudes notables. Aunque el método de Kucera no no sea tan útil, en lo que respecta a la **lingüística histórica**, como la solución adoptada por nosotros, bien merece cierta atención y por ello le dedico una sección de este capítulo. Una virtud de las virtudes del método de Kucera es la forma en que hace compara los fonemas de una y otra lengua: teniendo en cuenta la posición dentro de la sílaba, que también podía introducirse en mi cálculo de C_{fon}. Así pues con las precisiones sugeridas por el método de Kucera me dispuse a rehacer los cálculos.

Los resultados obtenidos se resumen en la siguiente tabla donde con el propósito de comparación se añade el resultado obtenido para una lengua no indoeuropea, el vasco (*euskera*). La primera columna da el valor

Tabla 5.4: Índices de correlación fonética C_{fon} de diversas lenguas con respecto al indoeuropeo clásico.

Lengua	*Índice global* $\boldsymbol{C_{fon}}$	*Vocales* $\boldsymbol{C_{fon}}$ (sólo vocales)	*Oclusivas* $\boldsymbol{C_{fon}}$ (sólo oclus.)	*Sonantes* $\boldsymbol{C_{fon}}$ (líq. y nasal.)	
Sánscrito	0,9319	0,8456	0,8732	0,8935	S|I
	0,8719	0,5955	0,9459	0,8368	I|S
	0,9019	**0,7206**	**0,9096**	**0,8652**	***Promedio***
Griego	0,8828	0,8630	0,8921	0,9207	G|I
	0,8472	0,8731	0,8076	0,8499	I|G
	0,8650	**0,8680**	**0,8498**	**0,8853**	***Promedio***
Latín	0,8718	0,8251	0,9232	0,8123	L|I
	0,8514	0,7649	0,7882	0,7853	I|L
	0,8616	**0,7950**	**0,8557**	**0,7988**	***Promedio***
Baltoeslavo	0,8462	0,7458	0,9190	0,8651	B|I
	0,8168	0,6162	0,7054	0,8270	I|B
	0,8315	**0,8122**	**0,8122**	**0,8461**	***Promedio***
Germánico	0,8106	0,6085	0,8979	0,9358	G|I
	0,7950	0,5967	0,8128	0,8877	I|G
	0,8028	**0,6026**	**0,8553**	**0,9117**	***Promedio***
Euskera	0,2976	0,0768	—	—	E|I
	0,2907	0,0605	—	—	I|E
	0, 2942	**0,0687**	—	—	***Promedio***

$H_X+H_Y \neq H_{XY}$

.....

nótese que normalmente para un lengua indoeuropea L C_{fon}(L|I) es mayor que C_{fon}(I|L), lo que significa que en general es más fácilmente predictible la ocurrencia de un fonema en la lengua L si conocemos la correspondiente palabra indoeuropea, que no al revés. Esta aparente coincidencia podría tener que ver con el hecho de que siempre es la lengua L la que deriva del indoeuropeo y no al revés.

[33] El grupo germánico nos es mal conocido en época antigua; los primeros textos proceden del gótico y datan del siglo IV. Por esta razón, y siempre que es posible, se escoge la forma gótica otras veces nos vemos obligados a tomar otra forma antigua por ejemplo del alto alemán antiguo, del anglosajón antiguo y ocasionalmente del inglés o alemán modernos.

[34] Por la misma razón que para el grupo germánico, nos vemos obligados a considerar simultáneamente diversas lenguas. Normalmente la forma escogida procede del lituano que es la lengua que presenta mayor arcaísmo, cuando esta falta o no está disponible se toma del antiguo eslavo, del letón, del antiguo prusiano y excepcionalmente del ruso.

[35] Kucera, H. (1962): "Statistical determination of isotopy", *Proceedings 9th int. Congress of Linguistics*.

5.6. Refinamientos del método y cuantificación mediante rasgos distintivos.

La medida de la correlación esbozada al final de la sección anterior, aunque en ciertos casos puede ser adecuada, adolece de ciertos problemas. Algunos de ellos como el **problema de la similitud fonética** y el **reemplazo en las formas léxicas** pueden eliminarse completamente introduciendo elementos adicionales. Otros como la extensión del método a **lenguas no emparentadas genéticamente** pueden resolverse tan sólo parcialmente.

El problema de la similitud fonética.

El método expuesto en la sección anterior tan sólo computa la regularidad de los cambios fonéticos, sin importar la semejanza de los sonidos implicados. Sin embargo, nosotros deseamos una medida efectiva de la similaridad entre dos lenguas [y no una mera estadística de lo predecible que son los valores de una variable aleatoria conocidos los de otra variable]. Por ejemplo, la **k-* indoeuropea evolucionada en las lenguas germánicas prácticamente siempre a *h-*, mientras que en latín siempre se conserva como *k-*. El latín está más cercano al indoeuropeo que las lenguas germánicas, por lo que a los sonidos velares se refiere. Es deseable por tanto introducir algo en nuestro método que de cuenta de que no es lo mismo una evolución $k > k$ que $k > h$, aunque ambas se den con idéntica regularidad estadística. El problema se resuelve si en lugar de examinar probabilidades conjuntas para fonemas particulares, examinamos probabilidades de conservación o **retención de rasgos fonéticos** particulares. Así por ejemplo en latín el rasgo distinctivo [+ oclusivo] siempre se conserva en posición inicial, mientras que en germánico puede dar lugar a [+ oclusivo] (para oclusivas sonoras) o [+ fricativo] (para oclusivas sordas). Si examinamos probabilidades de evolución de grupos de rasgos fonéticos distinctivos, conceptualmente la medida no cambia; sin embargo, cuantitativamente la medida es más precisa ya que tendría en cuenta, por ejemplo, que el cambio de *dh-* indoeuropea a *th-* en griego (indoeuropeo **dher* > gr. *ther* 'animal salvaje') más moderado[36] que el cambio de *dh-* indoeuropea a *f-* latina (indoeuropeo **dher* > lat. *fera* 'animal salvaje') o que el cambio de *dh-* indoeuropea a *d-* germánica (indoeuropeo **dher* > ingl. *deer*).

El problema de la exactitud en la correspondencia léxica.

Ya hemos dicho que para examinar la correlación fonética C_{fon} entre dos lenguas debe escogerse una lista de términos análogos suficientemente extensa y representativa. Sin embargo, hay un problema técnico; como muy bien saben los traductores no existe una relación biunívoca entre el vocabulario[37] de dos lenguas A y B: a veces una única forma léxica de la lengua A equivale, según el contexto, a dos o más formas léxicas en la lengua B; otras veces entre dos formas léxicas más o menos equivalentes en las lenguas A y B hay algún matiz semántico presente en una lengua y ausente en la otra; etc. Puede resultar que dos términos análogos W_A y W_B, tan sólo presenten coincidencia de significado en ciertos contextos.

Podemos corregir esta deficiencia de la siguiente manera, dividamos el campo semántico cubierto por W_A en dos partes: aquellas situaciones en las que su significado coincide con W_B y aquellas situaciones en las que su significado coincide con W_B (y análogamente para W_B). Si ahora introducimos las probabilidades de uso del vocablo W_A, en una y otra situación, podremos modificar convenientemente la distribución de probabilidad conjunta y dar una medida correcta que tenga en cuenta la inexactitud en la correspondencia léxica. El detalle técnico, no obstante, puede ser complicado, aunque para listas de términos básicos, culturalmente no marcados, en general el problema es poco grave ya que se trata de realidades concretas.

El problema de la redundancia y los hechos de distribución

Si se pretende un mejor ajuste de la medida que corrija la redundancia y los hechos de distribución (como por ejemplo la diferente evolución fonética según los fonemas vecinos) puede

[36] Ya que se conservan un mayor número de rasgos distintivos.

[37] Matemáticamente un vocabulario V_1 puede concebirse como una aplicación o correspondencia del conjunto de todos los campos semánticos posibles al conjunto de las secuencias

reducirse un pequeño refinamiento que corrige este error. El refinamiento consiste en estudiar probabilidades de ocurrencia no de fonemas individuales sino de grupos de fonemas[38].

Tomemos un caso concreto, un sonido [k] en latín, como en *cena* y *cantare*, se corresponde en castellano en proporciones similares con [k], como en *cantar*, y con [θ], como en *cena*, (y en proporciones menores con [ʧ]) por esta razón la imprevisibilidad será apreciable. En cambio si consideramos grupos de dos fonemas encontramos unas correspondencias precisas, por ejemplo, [kt] se corresponde casi exclusivamente con [ʧ] *noctem* > *noche*, [ke] con [θe], [ki] con [θi], [ka] con [ka], etc. por lo que la imprevisibilidad para grupos de dos consonantes es considerablemente menor. A medida que consideramos grupos más largos de fonemas la imprevisibilidad disminuye, es decir, vemos que la estructura global del latín explica mejor la estructura de las palabras castellanas que las simples correspondencias fonema a fonema.

Generalización a los otros niveles.

La generalización de nuestra medida de la información preservada o correlación a nivel **morfosintáctico** y **léxico-semántico** requiere de la lingüística estructural y de su caracterización de las estructuras mediante rasgos distinctivos. De esta manera, estos podrían utilizarse de manera completamente análoga a como se utilizan los rasgos fonéticos en el caso presentado.

Supongamos que pretendemos definir una medida de **correlación morfosintáctica** $C_{\text{mor-sin}}$. Podríamos empezar seleccionando una lista suficientemente larga y representativa de oraciones que muestren la estructura elemental de una lengua A y sus equivalentes en B. Tal lista debería estar formada por oraciones elementales que mostrasen entre otras: la flexión nominal y verbal, el orden de las palabras en la oración, las reglas de concordancia sintáctica, el tratamiento de las oraciones copulativas, y las oraciones de relativo. Seguidamente deberíamos descomponer estas oraciones en morfemas o en sus rasgos morfemáticos distinctivos. Cada morfema o agrupación de rasgos distinctivos se trata formalmente como un símbolo abstracto del que se calculan las probabilidades de ocurrencia y la probabilidad condicionada entre la aparición de un determinado elemento en la lengua A y sus elementos análogos en la lengua B, similarmente a como se hizo al analizar la correspondencia entre fonemas entre dos lenguas. Mediante estas dos distribuciones de probabilidad se evalúan I_{A} I_{B} y $I_{\text{A|B}}$ correspondientes a dichas distribuciones de elementos morfosintácticos y se calcula $C_{\text{mor-sin}}(\text{A}|\text{B})$ como[39]:

$$C_{\text{mor-sin}}(\text{A}|\text{B}) = 1 - \frac{I_{\text{A/B}}}{I_{\text{B}}} \qquad (1 \geq C_{\text{fon}} \geq 0)$$

Aquí el procedimiento es mucho más delicado que en el cálculo de C_{fon}, ya el número calculado dependerá enormemente de la muestra escogida y de lo representativa que ésta sea, por esto se exige que la lista sea larga ya que esto en principio garantiza mejores resultados. En mi opinión, el problema más grave que se presenta aquí es como decidir de una manera objetiva si una lista es representativa o no[40]. Para una medida de la **correlación léxico-semántica** $C_{\text{lex-sem}}$ procederíamos de modo análogo sólo que este caso es menos evidente escoger los símbolos entre los cuales calcular su correlación. Por otra, parte parece importante aquí escoger una lista en la que los diversos rasgos semánticos distinctivos aparezcan con la misma frecuencia que estos aparecen en la lengua común, porque de otra manera los datos correspondientes a I_{A} e I_{B} aparecerían muy falseados.

El problema del reemplazo de formas léxicas

Este problema tiene que ver con la desaparición de formas léxicas que caen en desuso y su reemplazo por otras nuevas, que nada tienen que ver fonéticamente con las primeras. Por ejemplo, derivadas de la palabra indoeuropea **rktos* 'oso' encontramos en sánscrito *rksas*, en griego *arktos* y en latín *ursus*, sin embargo en germánico y en eslavo encontramos ger. **bɛr*, esl. *medvedi*,

[38]Téngase en en cuenta que en la teoría expuesta la naturaleza exacta de los signos f_i no interviene en los resultados y nada nos impide interpretarlos como fonemas individuales o bien como grupos de fonemas

[39]Nótese que en general $C_{\text{mor-sin}}(\text{A}|\text{B})$ no tiene porqué ser igual a $C_{\text{mor-sin}}(\text{B}|\text{A})$.

[40]A parte de mostrar la estructura de lengua una lista debería tener en cuenta la probabilidad de uso en la lengua común.

respectivamente que nada tienen que ver con la raíz ***rktos***. En estas lenguas la raíz original ha sido reemplazada por una nuevas formas etimológicamente no relacionadas. Por tanto, una lista de términos análogos para medir la correlación entre por ejemplo, el latín y una lengua germánica, resultará problemática si en ella aparecen las palabras para 'oso'. Sin embargo podemos resolver este problema si introducimos una complicación adicional.

[::: -esbozar la solución]

Generalización a lenguas no emparentadas.

Generalizar la correlación fonético-fonológica a dos lenguas no relacionadas genéticamente puede resultar muy problemático. Por ejemplo, si comparamos euskera y latín es relativamente fácil hacer una asignación fonema a fonema entre palabras como *bake* (eus.) / *pacem* (lat.), o *pekatari* (eus.) / *pecatorem* (lat.). Sin embargo, en palabras como 'agua': *ur* (eus.) / *aqua* (lat.), 'dulce': *geza* (eus.) / *dulcis* (lat.), 'paloma': *uso* (eus.) / *columba* (lat.) existen múltiples posibilidades de asignación. Una posibilidad razonable es hacer la **asignación por sílabas** consistente en descomponer ambas listas de términos en sílabas y nombrar los fonemas de la misma en la forma:

$\ldots\text{-}I_3\text{-}I_2\text{-}I_1\text{-}N\text{-}C_1\text{-}C_2\text{-}C_3\text{-}\ldots$

dónde N es el fonema o grupo de fonemas que forma el núcleo de la sílaba (usualmente una vocal, un dipotongo o una fonema sonántico), I_1 es el fonema inmeidatamente precedente al núcleo de la sílaba (que puede estar ausente, si la sílaba empieza por vocal por ejemplo), I_2 es el fonema que precede a I_1 (que puede existir o no), análogamente C_1 es el fonema inmediatamente posterior al núcleo de la sílaba (que puede estar ausente como cuando la sílaba es abierta); y así sucesivamente. Por ejemplo la asignación por sílabas aplicadas a los pares anteriores de palabras tomadas del latín y el euskera sería:

(1)	*ur* - $\varnothing_s\varnothing_s$	*ge*$\varnothing_f$ - *za*$\varnothing_f$	$\varnothing_f$*u*-*so*$\varnothing_f$-$\varnothing_s\varnothing_s$	*ba* - *ke*$\varnothing_f$	*pe* - *ka* - *ta* - *ri*$\varnothing_f$
	a$\varnothing_f$ - *qua*	*dul* - *cis*	*co* - *lum* - *ba*	*pa* - *cem*	*pe* - *ca* - *to* - *rem*
	'agua'	'dulce'	'paloma'	'paz'	'pecador'

en las que se introducen los fonemas ficticios $\varnothing_f$ (fonema ausente) y $\varnothing_s$ (fonema correspondiente a sílaba ausente), que dan cuenta de las discrepancias en la diferente longitud de sílabas y palabras. A continuación se evalúan las probabilidades en las correspondencias obtenidas (en el ejemplo anterior: la *a* latina se corresponde[41] según el caso con *a*, *u*, o $\varnothing_s$ en euskera).

Como el número de fonemas que pueden actuar como centro de sílaba es bastante limitado en todas las lenguas, este procedimiento de asignación tiende a dar valores moderados de C_{fon} porque los fonemas nucleares se corresponden entre sí y los no nucleares entre sí, esto da como resultado un valor de C_{fon} de entre 0,20 y 0,30. Siendo una medida más adecuada el índice C_{fon} calculado a de las probabilidades de correspondencias que se dan únicamente entre fonemas no nucleares de una y otra lengua, este valor raramente supera 0,10.

5.7. El método de Kucera

Es instructivo presentar aquí el método de Kucera para la evaluación del **índice de isotopía**, que pretende ser una medida de la similitud fonológica de dos lenguas, emparentadas o no. De hecho hay bastantes similitudes entre el índice de isotopía de dos lenguas, designado mediante Is_{AB}, y el índice de correlación fonética $C_{fon,AB}$:

-Ambos índices toman **valores entre 0 y 1**.
-Ambos índices se basan en las **probabilidades de ocurrencia** de fonemas.

[41]Naturalmente estas correspondencias se dan al azar, y es esperar que un análisis probabilístico de las mismas revele que no existe ninguna correlación importante entre ambas lenguas. Aunque siempre habra una correlación marginal correspondiente a préstamos léxicos latinos como *bake* o *pekatari*.

- Si se **compara una lengua consigo misma** ambos índices valen 1, es decir $C_{fon,AA} = 1$ y $Is_{AA} = 1$. Aunque aquí hay una diferencia cualitativa el cálculo de $C_{fon,AA}$ da 1 de manera determinista con independencia del tamaño de la muestra, mientras que el cálculo de I_{AA} si se toma una muestra arbitrariamente grande, pudiéndose dar computaciones de este índice menores que 1 si la muestra no tiene el tamaño suficiente.

Sin embargo, la filosofía es diferente, el índice de isotopía tiene una inspiración sincrónica mientras que el índice de correlación fonológica tiene inspiración diacrónica. Los tres conceptos fundamentales en el método de Kucera son posición o distribución, sílaba fonológica e isotopía, que se definen a continuación. Por **posición o distribución** se entiende el lugar del fonema con respecto a sus fonemas vecinos. Por **sílaba fonológica**, siguiendo a Haugen, se entiende "la unidad fonémica secuencial más pequeña". Una sílaba fonológica tiene la siguiente estructura: inicio, núcleo y final; estos son sus constituyentes y están formados por grupos de fonemas. El **núcleo** de la sílaba el mínimo irreducible de ésta y está presente en cualquier sílaba, el resto de la sílaba constuyen los márgenes (inicio y fin). El **inicio** de la sílaba es la porción de la misma comprendida entre la separación silábica anterior y el núcleo de la sílaba, análogamente el **final** de la sílaba es lo que sigue desde el núcleo hasta la siguiente separación silábica.

Con estas definiciones previas Kucera se dispuso a calcular probabilidades de ocurrencia de fonemas en cada una de estas posiciones silábicas utilizando como material básico diversos textos en ruso y en checo. Por ejemplo, por lo que respecta al inicio de sílaba tanto en ruso como en checo este puede consistir en uno, dos, tres o cuatro fonemas o bien carecer de inicio y empezar directamente con el núcleo silábico. Estos cuatro diferentes tipos de inicios se llaman se llaman I_1, I_2, I_3 y I_4. Análogamente el final de sílaba puede ser de tipo F_1, F_2, F_3 y F_4 según el número de fonemas que lo formen, de hecho el tipo F_4 existe en ruso pero no en checo. De acuerdo con esto una sílaba tiene la forma Inicio + Núcleo + Final dónde:

Inicio	Núcleo	Final
I_{1A}	N	F_{1A}
I_{2A} I_{2B}		F_{2A} F_{2B}
I_{3A} I_{3B} I_{3AC}		F_{3A} F_{3B} F_{3C}
I_{4A} I_{4B} I_{4C} I_{4D}		F_{4A} F_{4B} F_{4C} F_{4D}

Así I_{3B} denota un fonema que aparece en segunda posición en una sílaba cuyo inicio está constituido por dos fonemas, F_{1A} sería el único fonema final de un sílaba, etc. Por otra parte, tanto en ruso como en checo los fonemas que aparecen tanto al inico como al final son las consonantes más el fonema /j/; por el contrario en ruso el núcleo está formado siempre por una vocal única, mientras que en checo en núcleo puede ser una vocal , un diptongo o bien una de las sonantes /r/. /l/ o /m/. Dos fonemas que pueden aparecer en posiciones silábicas idénticas se llaman **isotópicos**. Así el índice de isotopía es una medida estadística basada en la diferente porbabilidad de ocurrencia de dos fonemas isotópicos, cuanto más diferentes sean estas diferencias entre las probabilidades de ocurrencia tanto más bajo será el índice de isotopía.

5.8. Cuantificando la redundancia en el lenguaje.

Cuando en una lengua escrita cierta letra aparece en un texto, entonces la probabilidad a priori de que le siga otra letra dada no es la misma para todas las letras. Por ejemplo, en las lenguas del mundo el grupo inicial *mr-* es muy escaso, si una palabra empieza por *m-* difícilmente la siguiente letra será *r* ([r] o [ɾ]). En inglés escrito, por ejemplo, si una palabra empieza por t- es mucho más probable que le siga *h* que no, por ejemplo, *n* ya que el grupo *th-* es relativamente frecuente. Similarmente dada la secuencia castellana *ció* la probabilidad de que la siguiente letra sea *n* es extremadamente alta, hasta el punto de que es casi redundante la información consistente en decir que la siguiente letra es efectivamente *n*. Las correlaciones como las descritas, en la que la ocurrencia de un cierto signo en el código es inferible con cierta probabilidad a partir de los otros signos (inferible a partir del contexto) es lo que llamamos **redundancia**.

Todas las lenguas naturales presentan un grado considerable de redundancia. Ésta redundancia tiene por objeto **asegurar la comunicación**, sobrecaracterizando los mensajes. Si en la emisión o percepción de un mensaje se produce algún error (*e.g.* confusión o eliminación de algún fonema), el mecanismo de redundancia evita los **malentendidos** y **ambigüedades** ya que la misma información se halla de alguna manera sobrexpresada[42]. Es lo que llamamos, entender algo "por el contexto". En el siguiente ejemplo se han omitido las vocales pero aún así el mensaje sigue siendo inteligible:

(2) *L MYR D GNT PD LR ST MNSJ SN PRBLMS* [43]

Naturalmente las lenguas naturales representan un equilibrio entre los **códigos superabundantes** (en los que la probabilidad de malentendido o ambigüedad es prácticamente nula) y los **códigos mínimos** (que utilizan el mínimo número posible de signos para expresar la información, pero en los que prácticamente cualquier error ocasiona confusiones).

Aplicaciones

El estudio de la redundancia tiene dos aplicaciones principales. Dentro de la lingüística permite cierta predicción sobre la plausibilidad de ciertos cambios (o errores) lingüísticos. Dentro de la **criptografía** general pemite diseñar códigos más eficientes (de mayor economía de signos).

Para ilustrar el diseño de un código más eficiente consideremos un modelo de lenguaje que conste tan solo de cuatro signos s_1, s_2, s_3 y s_4 con probabilidades de ocurrencia iguales a p_1=🗁, p_2=🗏, p_3=🖮 y p_4=🖮. Si tratamos de codificar dicho lenguaje mediante un código binario la posibilidad más sencilla sería hacer, por ejemplo, $s_1 \rightarrow 00$, $s_2 \rightarrow 01$, $s_3 \rightarrow 10$ y $s_4 \rightarrow 11$. Este código requiere dos dígitos binarios o bits[44] por signo ($I_{\text{símbolo}}= 2$) de mensaje. Haciendo uso de las probabilidades de aparición es sencillo establecer el código más eficiente es: $s_1 \rightarrow 0$, $s_2 \rightarrow 10$, $s_3 \rightarrow 110$ y $s_4 \rightarrow 111$ [Es inmediato comprobar que cualquier a partir de una cualquier secuencia de 0 y 1 que codifique un cierto mensaje es posible recobrar la secuencia original de signos s_1, s_2, s_3 y s_4]. Por otra parte, el número medio de bits requeridos por signo es menor: 🗁·1+🗏·2+🖮·3+🖮·3 = 1,75 < 2. Este código es mínimo ya que el número medio de signos coincide con $H = -(p_1 \ln p_1 + p_2 \ln p_2 + p_3 \ln p_3 + p_4 \ln p_4) = 1{,}75$.

Esta idea básica, de usar códigos cortos para signos (o letras) frecuentes y códigos largos para signos poco frecuentes, está presente en el **código Morse**. Igualmente es más probable que de dos palabras W_A y W_B que designan una misma realidad material en dos lenguas A y B, sea más breve aquella correspondiente a la lengua o cultura en la que dicha realidad sea más frecuente. Por ejemplo, David Crystal menciona el hecho de que en lengua pintupí[45] existen palabras sencillas como *nyarrkalpa*, *pulpa* o *katarta* para las que no existen traducciones simples y deben traducirse por expresiones tan largas como 'madriguera de un animal pequeño', 'agujero hecho por un conejo' y 'agujero que deja una iguana cuando rompe la superfície después de la hibernación'. Todos estos conceptos son más habituales para los hablantes del pintupí que para nosotros, por eso han desarrollado palabras simples para nombrarlos. Inversamente los pintupís han utilizar largos giros lingüísticos para designar conceptos habituales para nosotros como *periodista* o *diccionario*.

Cuantificación de la redundancia.

Para cuantificar la redundancia en una lengua vamos a recurrir al concepto de probabilidad condicionada. Esta herramienta matemática nos permite analizar correctamente la correlación entre la aparición de un fonema en una palabra y la ocurrencia de otros fonemas en la misma palabra. Para cada par de letras definamos dos variables aleatorias X e Y, la primera letra será el valor de X y la segunda el valor de Y, es evidente que los valores potenciales a priori de X e Y son todo el alfabeto. En estas condiciones **la información condicional** viene dada por signo la ecuación [5]. Esta correlación definida para pares de letras puede extenderse fácilmente a grupos o secuencias de

[42]Técnicamente deberíamos decir codificada o implementada varias veces. Es como, por expresarlo metafóricamente, si se nos pinchara una rueda del coche y tuviéramos otra de recambio.
[43] *La MaYoRía De GeNTe PueDe LeeR eSTe MeNSaJe SiN PRoBLeMaS*.
[44]*bit* < *binary digit* 'dígito binario' (0 o 1).
[45]Lengua aborigen australiana; Crystal, D. (ed.) (1994): *Enciclopedia del lenguaje*, Taurus, Madrid.

letras (e.g. *-ció-*, *th-*, *mr-*, etc.) sin más que definir las variables aleatorias adecuadamente. Denotemos una secuencia de N-1 signos (letras o fonemas) mediante[46] $b_i(N-1)$ y llamemos $p_{(N-1)}(b_i(N-1))$ a su probabilidad de ocurrencia, esta es nuestra primera variable aleatoria. Nosotros deseamos discutir la probabilidad que esta secuencia sea seguida por un cierto signo o letra, formando una secuencia de N signos, denotaremos la probabilidad conjunta de una secuencia $b_i(N-1)$ seguida del j-ésimo signo por $p_{(N-1),X}(b_i(N-1), j)$, si llamamos F_N a la información condicional asociada tenemos que, acorde a la fórmula [5] *mutatis mutandis*:

$$F_N = -\sum_{i,j} p_{(N-1),X}(b_i(N-1), j) \log_2 p_{X|(N-1)}(j|b_i(N-1))$$

(Dónde la suma sobre i se extiende sobre todas las posibles secuencias de N-1 letras, y la suma sobre j sobre el alfabeto). Si ahora consideramos el límite cuando N crece indefinidamente tendremos una medida de toda la estructura de la lengua sobre el signo j. La cantidad límite F_{lim} se define el valor al que se aproxima la cantidad F_N cuando N crece, es decir:

$$F_{lim} = \lim_{N\to\infty} F_N$$

Como esta ya es una medida correcta de la información por signo (ver 5.4), podemos cuantificar la redundancia *R* como:

[6] $$R = 1 - (F_{lim}/F_0)$$

siendo F_0 la imprevisibilidad suponiendo equiprobabilidad a priori de todas las letras. Las primeras F_N computadas para el inglés escrito son:

F_0= 4,76 bits	(equiprobabilidad de todas las letras: 27 signos, incluyendo ‘ ’)
F_1= 4,03 bits	(usando las probabilidades individuales de cada letra)
F_2= 3,32 bits	(usando datos sobre pares de letras)
F_3= 3,1 bits	(usando datos sobre secuencias de tres letras)

Para N>3 existen pocos datos, incluso para el inglés escrito, por lo que F_{lim} es mal conocido; medidas alternativas[47] a la computación directa de largos textos mediante equipos informáticos sugieren que F_{lim} estaría próximo a 1,4 bits. Por lo que la redundancia estimada para el inglés escrito está en torno a $R = 0{,}8$ (es decir, el inglés escrito es redundante en un 80%, deberíamos eliminar un 80% de los signos para convertirlo en un código mínimo).

5.9. Ley de Zipf-Mandelbrot sobre la frecuencia de las palabras.

En 1949 el psicolingüista estadounidense G. K. Zipf[48] propuso una fórmula empírica para la frecuencia de aparición de las palabras en una lengua. Supongamos primero que ordenamos todas las palabras del diccionario de más frecuente a menos frecuente; al orden que ocupa una palabra en esta clasificación le llamaremos **rango** de la palabra. De esta manera, la ley de Zipf relaciona el rango *n* de una palabra con su probabilidad (= frecuencia) de aparición p_n en un mensaje cualquiera:

[7] $$p_n = P/n$$

siendo *P* una constante relacionada con el número total de palabras usadas y que Zipf evaluó aproximadamente en $P = 0{,}1$[, lo que corresponde a un número de palabras $N = 8727$, ya que la suma de todas las probabilidades deber ser igual a 1]. La siguiente gráfica[49] muestra el ajuste de los resultados experimentales a la fórmula de Zipf.

:::

[46] Ahora i tomará tantos valores como combinaciones de N-1 puedan formarse.
[47] Shannon, C. E. (1951): *Bell System Techonology Journal*, 30, pp. 54-58.
[48] Zipf, G. K. (1949): *Human Behaviour and the Principle of Least Effort*, Addison-Wesley, Cambride, Massachusetts.
[49] Brioullin, L. (1962): *Science and Information theory*, Academic Press, Nueva York.

La *fórmula empírica* de Zipf funciona razonablemente bien, tal vez su único problema es que en sí misma no proporciona explicación teórica de porqué funciona. Sin embargo, en 1960 se produjo un avance importante, cuando B. Mandelbrot dedujo *teóricamente* otra fórmula un poco más general que incluía como caso particular a la de Zipf. Mandelbrot buscó de todas las distribuciones posibles aquella que optimizaba el intercambio de información, es decir, de todas aquellas con un mismo **tiempo de articulación** encontró la que daba más información[, más adelante discutiremos la deducción de Mandelbrot]. La fórmula encontrada por Mandelbrot es:

[8] $$p_n = P(n+B)^{-\gamma}$$

(donde P, B y γ son constantes características de cada lengua). La fórmula de Zipf [7] se obtiene a partir de la de Mandelbrot se toma $B = 0$, y $\gamma = 1$. La fórmula de Mandelbrot [8] no sólo se ajusta mejor a los datos experimentales, sino que además tiene un interés para la lingüística mucho mayor que la de Zipf; primero porque su deducción proporciona una explicación teórica de porque se ajusta a la realidad y segundo porque las constantes P, B y γ pueden interpretarse y proporcionar información adicional sobre la lengua.

En primer lugar P está relacionado con el **número total de palabras** R como en la fórmula de Zipf, conocido P puede evaluarse R, y viceversa conocido R puede evaluarse P, la ecuación que relaciona ambos parámetros se obtiene, del hecho de que la suma de todas las probabilidades debe ser igual a 1:

$$\sum_{n=1}^{n=R} p_n = p_1 + p_2 + \ldots + p_R = 1 \qquad \Rightarrow \qquad P = 1 \Big/ \sum_{n=1}^{n=R} (n+B)^{-\gamma}$$

El otro parámetro interesante es γ. Para la mayoría de lenguas se ha encontrado $1{,}2 \geq \gamma \geq 1$, aunque en casos excepcionales como en el caso del habla infantil puede alcanzar el valor $1{,}6$. Es decir, los niños tienden a comunicar sus ideas haciendo uso de palabras frecuentes con mayor frecuencia que los adultos, puesto que $\gamma = 1{,}6$ favorece el uso de palabras frecuentes que generalmente son también las más cortas. Por tanto, de alguna manera γ es una medida indirecta de la diversificación del vocabulario.

Las siguientes tablas ilustran datos experimentales referentes a diferentes lenguas y su comparación con la ley de Zipf-Mandelbrot:

[:::]

Discusión de la fórmula de Zipf-Mandelbrot

Presentaremos una deducción de la fórmula de Mandelbrot (a la que de ahora en adelante nos referiremos como Ley de Zipf-Mandelbrot) basada en el formalismo de la teoría de la información.

Preguntémonos en primer lugar por la distribución de frecuencia fonemática óptima, es decir, aquélla que aporta una mayor información, para un mensaje de longitud dada. Supongamos que nuestro lenguaje (una lengua, un texto escrito, una muestra del habla de un hablante particular) consta de m símbolos (según el caso: fonemas, letras o sonidos) y que cada símbolo s_j (j=1, ..., m) tiene una duración o coste generalizado t_j. Consideremos por otra parte, todos los mensajes de duración T (más generalmente, puede pensarse en T como un coste generalizado); de todos estos queremos encontrar aquél con una distribución de símbolos que maximice la cantidad de información. Digamos también que el mensaje que maximice la información conste de N símbolos: N_1 del tipo s_1, N_2 del tipo s_2, etc. y, por tanto, la probabilidad del símbolo j-ésimo vendrá dada por $p_j = N_j/N$. Entonces lo que pretendemos es maximizar la función H:

$$H_{mensaje} = N \cdot H(p_1,\ldots,p_m) = -N \sum_{j=1}^{j=m} p_j \ln p_j$$

sujeta a las dos siguientes condiciones:

$$[*] \quad T = \sum_{j=1}^{j=m} N_j t_j = N_1 t_1 + \ldots + N_m t_m$$

$$[**] \quad \sum_{j=1}^{j=m} \frac{N_j}{N} = \frac{N_1 + \ldots + N_m}{N} \qquad \text{es decir:} \quad \sum_{j=1}^{j=m} p_j = 1, \quad \text{ya que} \quad p_j = \frac{N_j}{N}$$

[La condición [*] expresa que la duración o coste generalizado total, es igual a la suma de las duraciones o costes generalizados de todos los símbolos que componen el mensaje. La condición [**] expresa que la suma de todas las probabilidades es igual a 1.]

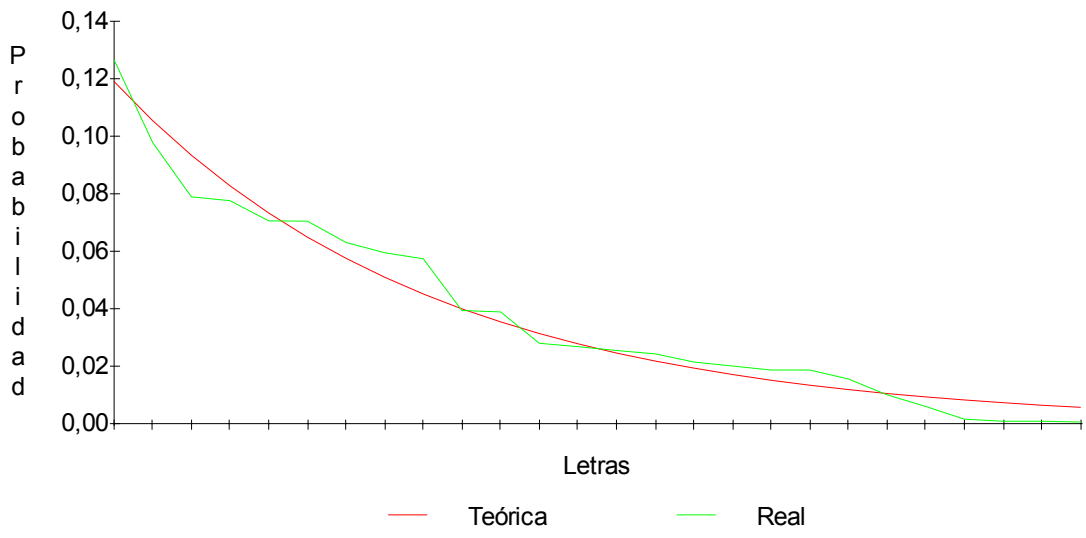
El problema así planteado, es formalmente idéntico al planteado en física de determinar el estado macroscópico de un sistema de moléculas conocidos los niveles energía. La identificación de las cantidades que aparecen en el problema físico y en el problema aquí planteado van como sigue: el estado macroscópico del sistema desempeña el mismo papel que el mensaje, la energía total del sistema E es el análogo de T en nuestro caso; y los costes generalizados t_j corresponderían en el problema físico a los diferentes niveles energéticos moleculares ε_j. Más adelante seguiremos con otros aspectos de esta analogía que nos permitirán encontrar magnitudes útiles para la lingüística.

Técnicamente el problema se resuelve mediante el método de los multiplicadores de Lagrange, que lleva a unas ecuaciones cuya solución es:

[9] $p_j = \exp(-\beta_1 t_j)$ dónde β_1 es una constante tal que se satisface: $p_1 + \ldots + p_m = 1$

como ilustración de la validez empírica de esta fórmula se comparan las frecuencias de aparición de las diferentes letras de un texto en inglés con una distribución teórica, suponiendo que todos los t_j tienen aproximadamente el mismo coste generalizado más un factor lineal (es decir, $t_j = t_0 + \varepsilon j$, dónde ε es una cantidad pequeña):

Tabla 5.5: *Probabilidad de ocurrencia de las diferentes letras en un texto en lengua inglesa.*



Naturalmente un modelo que presuponga t_j diferentes puede reproducir la curva real de manera exacta, el interés del gráfico anterior es que viene a decir que la distribución real no está demasiado lejos de una distribución en que el coste generalizado de los símbolos tienen varía poco y de manera lineal. Por otra parte si la ortografía inglesa se ajustase más a una ortografía fonemática es posible que muchas de las discrepancias desaparecieran.

Se puede demostrar también, que el número de palabras N_p de duración o coste menor o igual que T viene dado por la expresión:

[10] $N_p = A \exp(-\beta T) - B$, o equivalentemente $\beta(t-t_0) = \ln (N_p+B)$

Siendo B, A, β y t_0 constantes. Con este resultado para la distribución de fonemas (símbolos) estamos en condiciones de deducir la ley de Zipf-Mandelbrot. Como la ecuación [9] también es aplicable cuando los símbolos son palabras enteras, puede escribirse para la palabra de rango n:

[9] $p_n = \exp(-\beta_1 t_n)$ siendo t_n el coste de esta palabra

De estas dos últimas ecuaciones puede eliminarse el coste t_n que no sabemos como determinar empíricamente[50] y podemos escribir: $-\beta_1 t_n = ln\ p_n = -\beta_1 t_o - (\beta_1/\beta) \ln (N_p+B)$ y de aquí:

$$p_n = \exp(-\beta_1 t_o)\ (N_p+B)^{-(\beta 1/\beta)}$$

que es la **ley de Zipf-Mandelbrot** [8] con $P = \exp(-\beta_1 t_o)$, $n = N_p$ y $\gamma = \beta_1/\beta$. Aparte de esta deducción de la ley de Zipf-Mandelbrot, hay una deducción geométrica debida a Mandelbrot[51], cuyo formalismo escapa a los intereses de este libro. Sin embargo, la deducción geométrica permite una interpretación más clara de los parámetros que aprecen en [8]. Así si m es el número de signos (letras, fonemas, etc.) del lenguaje considerado y L la longitud media de una palabra, Mandelbrot nos dice que en $p_n = P(n+B)^{-\gamma}$:

$$P = \left[\frac{2(m-1)}{m}\right]^{-\gamma} \frac{1}{1+L} \qquad \text{y} \qquad B = \frac{1}{m-1}$$

Estas fórmulas nos relacionan las constantes, P y B a parámetros fácilmente medibles como son m y L. De esta manera, podemos ver que la fórmula original de Zipf con $B = 0$ y $P = 0{,}1$ corresponde a una lengua con un gran número de signos $m = \infty$ y $L = 4$. Por otra parte el exponente γ, no tiene una interpretanción tan directa[52]. Por otra parte la deducción geométrica también establece que las palabras más frecuentes o de menor rango son también las más cortas, es decir, la constan de un menor número de símbolos (fonemas, letras, sonidos, etc.). De esta manera, puede interpretarse que la ley de Zipf-Mandelbrot está relacionando de alguna manera la probabilidad de uso de una palabra con su tiempo de articulación (o, más en general, con un coste fisiológico generalizado).

Otro hecho interesante de la ley de Zipf-Mandelbrot, que nadie ha advertido directamente hasta ahora[53], es que la ley de Zipf-Mandelbrot implica que ha de existir un equilibrio o relación entre el **número de formas léxicas distinctas** (el número de vocablos listables en un diccionario) y la **longitud media de las palabras** en una lengua. Entre dos lenguas con R_1 y R_2 formas léxicas ($R_1 < R_2$) es de esperar[54] que $\gamma_1 > \gamma_2$ y, por tanto, la probabilidad de encontrar palabras largas es mayor en R_2 que en R_1.

[50] Sería interesante determinar de alguna manera esta magnitud. En la analogía entre imprevisibilidad y entropía de la que hablamos al inicio del capítulo $1/\beta$ es el análogo de la temperatura. Conocidos los t_n podríamos estimar $1/\beta$; es plausible que esta magnitud sea tan importante para la lingüística como la tempuratura lo es para la física.

[51] Mandelbrot, B. (1977): *The fractal geometry of nature*.

[52] Según Mandelbrot $1/\gamma$ corresponde a la dimensión del árbol lexicográfico formado por todas las palabras de la lengua, que tendrá una estructura geométrica fractal (por lo que la dimensión puede no ser un número entero).

[53] Al menos por lo que, a mí, el autor de este libro, me consta.

[54] Si suponemos que el parámetro P es aproximadamente la misma para ambas lenguas.