Léxico poético, frecuencias, concordancias y estadística. Análisis de *Fêtes de la Patience* de Á. Rimbaud

Gabriel Mª JORDÀ LLITERAS
Ángel IGELMO GANZO
Universitat de les Illes Balears

En estas líneas pretendemos exponer los aspectos básicos de un proyecto de trabajo emprendido por el Área de Filología Francesa y la de Análisis Matemático y Estadística de la UIB. Nuestro objetivo es simple: utilizando la potencia y versatilidad de los actuales microprocesadores, elaborar un método de trabajo que nos permita facilitar a los estudiosos de la lengua o de la literatura documentos de trabajo, que ofrezcan análisis detallados del lenguaje de obras o textos poéticos.

Conocemos algunos de los excelentes trabajos realizados en este campo; permítasenos citar, entre otros, los de Sabido Rivero y de García Wiedemann, de la Universidad de Granada, o los análisis de P. Guiraud, de Finzi, de Elsa Dehennin o los del *Centre d'Étude du Vocabulaire Français* de la *Faculté des Lettres* de Besançon, llevados a cabo bajo la dirección de B. Quemada. A esta línea de investigación que, con la aparición de los primeros grandes ordenadores, comenzó a desarrollarse en la década de los sesenta, pretendemos contribuir con dos pequeñas aportaciones: la primera, conseguir aumentar la facilidad y rapidez, tanto en la realización de los análisis, como en la consulta o el estudio por parte del investigador, (para ello se presentan las notas introductorias, los resúmenes, cuadros y análisis estadísticos, el léxico base y el léxico general en soporte de papel, mientras que las concordancias se ofrecen en un disquete con un programa que permite su rápida consulta e impresión), la segunda, aplicar el cálculo de contraste de hipótesis estadísticas efectuadas sobre muestras y población.

Somos conscientes de que, aunque la lingüística estadística y computacional constituye una rama importante dentro del estudio del lenguaje, los trabajos realizados en este campo corren el riesgo de verse reducidos a simples inventarios de los que sólo se extraigan conclusiones vagas o tautológicas. En ningún momento debe olvidarse que así como el templo es la razón de ser de las piedras, es el poema quien da sentido a las palabras. Ahora bien, nuestro

Emilia Alonso et al (eds.), La lingüística francesa: gramática, historia, epistemología, Tomo I, Sevilla, 1996

objetivo concreto no estriba en la obtención de conclusiones, sino, simplemente, en el análisis de un corpus poético y en la elaboración de documentos de trabajo; sin embargo nos parece importante recordar algunos aspectos fundamentales del marco teórico de nuestra línea de investigación.

En sus *Essais de linguistique générale*, Jakobson nos recuerda que «un linguiste sourd à la fonction poétique, comme un spécialiste indifférent aux problèmes et ignorant des problèmes linguistiques sont d'ores et déjà l'un et l'autre de flagrants anachronismes» (1963: 248) y, siguiendo esta línea de pensamiento, P. Guiraud propone una lectura vertical, estructural, que reconstruya el campo estilístico de cada palabra a partir del conjunto de sus empleos, de la totalidad de las situaciones en las que se ve implicada. Para Guiraud, la noción de campo estilístico en lexicología surge cuando se considera que un léxico es una red de interrelaciones entre las palabras que lo constituyen, lo que, tradicionalmente, ya ha sido más o menos implícitamente admitido al estudiar algunas palabras clave en la obra de un autor.

Desde los años sesenta se considera toda obra literaria como un mensaje del cual hay que reconstruir el código, para ello se insiste en el carácter estructural de este código como un sistema en el que cada signo se define por su relación con los demás. «Un tel code est stylistique dans la mesure où il est propre à une oeuvre et à un auteur. Il tire alors son sens et sa fonction d'un système spécifique et qui constitue un écart par rapport au réseau de relations que ces mêmes mots entretiennent dans la langue commune» (1969: 86), y Eisa Dehennin, comentando a Jakobson, afirma que: «En parlant de certaines variétés stylistiques, Jakobson ne parle-t-il pas des sous-codes du code total? Sans pour autant constituer un «écart» (un idiolecte si l'on veut) l'oeuvre n'est-elle pas comme un code dans le code? Un sous-code non pas informe, non pas infini, mais cohérent, structuré même, puisqu'il a été sélectionné par un émetteur-créateur et combiné pour transmettre au receveur-lecteur le contenu d'un message, non moins cohérent et lourd de signification, et cela, sans aucun doute, selon le principe de l'équivalence linguistique et immanente, dont il a été question, mais aussi, nous semble-t-il, selon le principe d'une équivalence translinguistique interne et transcendante. Car une autre équivalence s'établit au niveau du sous-code entre le signifié global de l'oeuvre et ses signifiants «médiateurs» - pour employer un terme de R. Barthes. L'oeuvre n'est-elle pas un signe motivé?» (1969: 13). No creemos necesario insistir sobre el interés que los análisis de corpus establecidos por especialistas puede despertar al estudiar textos de una época o movimiento determinados, o, al abordar toda la obra de un autor, dividida por espacios cronológicos, considerar las preferencias por unas voces antes que por otras, así como los temas, cuáles se ven aumentados, y cuáles disminuidos a través del tiempo.

En nuestra opinión es también muy importante considerar todas las palabras gramaticales o semánticamente débiles; Guiraud nos advierte que «la linguistique moderne - structurale et fonctionnelle - montre l'originalité et les pouvoirs expressifs des formes, en apparence les plus neutres et les plus insignifiantes... on découvre alors que le choix des formes grammaticales détermine les modes de la communication et le sens intime de la «vision du monde» (1969: 139), y el Dr. Vicente Sabido Rivero afirma que «al linguista, al

Emilia Alonso et al (eds.), La lingüística francesa: gramática, historia, epistemología, Tomo I, Sevilla, 1996

filólogo en un sentido más amplio, le interesan todas y cada una de las palabras de un texto, no sólo las más significativas» (1992: 16).

Para llevar a cabo nuestro trabajo hemos utilizado un PC compatible IBM 486 DX4, 100 MH y 8 Megas de RAM, un escáner de mesa, con OCR profesional, algunas herramientas de los procesadores de texto WordStar y WordPerfect, DBase III Plus y Clipper 5.01, con los que hemos desarrollado una aplicación que nos permite obtener las concordancias, frecuencias y porcentajes; y, para aquilatar nuestro método, hemos optado por realizar el análisis de los cuatro poemas que componen las *Fêtes de la Patience* de A. Rimbaud (1984: 107-111), considerados por la crítica como una suerte de «mini-corpus» dentro de la poesía del poeta, por lo que, en principio, el lenguaje debe ser altamente homogéneo; las abreviaturas utilizadas han sido: BM: *Bannières de mai*, CH: *Chanson de la plus haute tour*, LE: *L'Éternité*, AO: *Âge d'or* y, para el conujunto total, FP: *Fêtes de la Patience*.

Las etapas básicas han sido las siguientes: obtención del índice de todas las palabras, codificación gramatical, marcaje del texto (desambiguación), codificación de la obra, obtención de concordancias, frecuencias y porcentajes sobre formas gramaticales, total de palabras y total de versos y, finalmente, análisis estadísticos. En las *Frecuencias y porcentajes* la primera cifra corresponde a la frecuencia absoluta, la segunda al porcentaje sobre el total. En la *Clasificación y distribución gramatical* las frecuencias y porcentajes están calculados sobre el total de formas diferentes y sobre el total de palabras distintas de cada uno de los poemas y de las FP en su globalidad, en el *Léxico general* sobre el total de versos de cada poema y el del conjunto de las FP en los que aparece la forma en cuestión. En los diferentes recuadros de la *Clasificación y distribución gramatical* se ofrecen el número de palabras, el número de formas, las categorías gramaticales con su correspondiente codificación y su frecuencia y porcentaje. En el *Léxico base* la cifra que sigue a cada una de las palabras remite a las diferentes formas del *Léxico general* con las que dicha palabra aparece dentro de las FP. En el *Léxico general* cada forma va seguida de su codificación gramatical, de su frecuencia/porcentaje en cada uno de los poemas y en el conjunto de las FP (con los cual podemos establecer un índice de similitud) y, finalmente, si se utiliza soporte de papel, del número de página en la que pueden encontrarse las respectivas concordancias; en caso contrario, si utilizamos soporte informático, aparece, si ha lugar, la «marca» de la forma. En las *Concordancias* el código que antecede a cada uno de los versos permite su rápida localización dentro de la obra codificada que se ofrece al inicio del trabajo.

Desde el punto de vista *estadístico*, el análisis lexico debe considerar las palabras individualmente, sin conexión (sintáctica) entre ellas. Es decir, cuando analizamos estadísticamente las palabras de un texto, debemos considerarlas como elementos individuales de un *referencial*, representado por el conjunto (en sentido matemático) de las palabras que componen dicho texto. Una imagen de esta situación sería como si hubiéramos metido todas las palabras del texto en una urna y contáramos el número de veces que cada palabra se extrae de la urna, lo que constituiría su frecuencia absoluta. Evidentemente, este conjunto de modalidades de la variable «palabra» es demasiado extenso.

Emilia Alonso et al (eds.), La lingüística francesa: gramática, historia, epistemología, Tomo I, Sevilla, 1996

El siguiente paso consiste en organizar estas formas agrupándolas en clases (en sentido estadístico), siendo el *agrupamiento natural* el que resulta de clasificar las palabras según su naturaleza gramatical. Para la codificación gramatical hemos seguido la adoptada por el *Centre d'Étude du Vocabulaire Français* de la *Faculté des Lettres* de Besançon, a la que se le han añadido dos «marcas»: + (interjecciones) y $ (otras categorías). Esta situación ha sido reflejada en los cuadros adjuntos: se han elegido doce agrupamientos, contándose los elementos de cada clase (frecuencia absoluta) y calculando los porcentajes, con lo que hemos obtenido lo que estadísticamente se llama una *distribución de frecuencias*. A partir de la distribución de frecuencias podemos llevar a cabo el análisis estadístico cuyas características fundamentales exponemos a continuación:

1. CONSIDERACIONES SOBRE LA POBLACIÓN Y EL MUESTREO:

Para realizar procesos de *inferencia estadística* es preciso concretar con precisión la *población de referencia* y el *proceso de muestreo*.

Por ejemplo, podemos considerar como población toda la obra literaria de un autor, o sólo una determinada obra, o un texto apócrifo; podemos tomar un estilo, o una época, o la creación de varios autores. Esta población de referencia deberá ser establecida por el investigador en función de sus intereses. Las técnicas estadísticas de *inferencia* nos permiten establecer el grado de similitud entre la lengua y estilo utilizados en los diferentes casos.

En muchos casos sólo se dispondrá de una parte del texto. Para que esta parte del texto pueda ser considerada como muestra deberá ser representativa de la población, por lo que es preciso definir lo que entendemos por *muestra*: entendemos por *muestra* cualquier parte del texto que tenga sentido sintáctico y que sea inteligible para el lector, es decir, que posea una estructura lógica.

2. REPRESENTACIÓN GRÁFICA:

Un tipo de diagrama útil sería el clásico histograma. Consideramos interesante emplear un gráfico de frecuencias acumuladas que indicaría el grado de concentración alrededor de ciertas clases, para ello se ordenarían las frecuencias de mayor a menor y representaríamos en *ordenadas* las frecuencias acumuladas, y en *abcisas* las clases (de 0 a 11). En estudios económicos esta gráfica se llama curva de *Lorentz*, aquí podemos llamarla *Curva de Concentración Lexicográfica (CCL)*; esta curva nos ofrece una información visual sobre el grado de coincidencia del estilo de dos o más textos.

Paralelamente a esta curva puede utilizarse el índice de *Gini*, que es el área comprendida entre la curva y la diagonal. Este índice nos mide el grado en que la curva se aparta de la diagonal. Este índice lo denominamos *Índice de Distribución Lexicográfica (IDL)*.

3. COMPROBACIONES ESTADÍSTICAS:

Proponemos utilizar las técnicas estadísticas de contraste o verificación de hipótesis sobre el parámetro de las proporciones. En todos los casos se utiliza

Emilia Alonso et al (eds.), La lingüística francesa: gramática, historia, epistemología, Tomo I, Sevilla, 1996

un *nivel de significación* del 5% (aunque otros valores corrientemente utilizados son los de 1% y 10%).

Hemos trabajado sobre los siguientes casos:

a) Verificación de una proporción con un valor prefijado:

Se trata de verificar si las proporciones que aparecen en un texto (considerado como muestra) son equivalentes a las proporciones obtenidas (conocidas) de un texto más amplio (considerado como población). Tomamos como hipótesis nula (verdadera) cada una de las proporciones de la población y contrastamos estos valores con los valores muestrales. Aceptar la hipótesis nula significa que las diferencias entre la muestra y las de la población son debidas al azar, es decir, que ambos textos presentan un mismo estilo. En caso contrario, rechazar la hipótesis nula, significa que no podemos considerar ambos estilos «equivalentes» o «compatibles».

El estadístico de contraste que se utiliza en estos casos corresponde a la fórmula nº 1. Si el valor del estadístico de contraste sale entre -1,96 y +1,96 se acepta la hipótesis nula. En caso contrario rechazamos dicha hipótesis, pero teniendo en cuenta que la probabilidad de rechazo, aunque sea cierta la hipótesis nula, es igual al nivel de significación (alfa = 0,05). Este proceso debe hacerse en cada proporción.

Hemos llevado a cabo esta verificación con los adjetivos calificativos (4), tomando como valor de referencia $P_0 = 0,1021$; para BM con el valor $P_m = 0,0588$ obtenemos el valor del estadístico de contraste $z = -1,8646$, por lo que se acepta la hipótesis nula; sin embargo, en el poema AO con $P_m = 0,1488$ obtenemos un valor de $z = 1,999$, por lo que la hipótesis nula debe ser rechazada.

b) Comparación de dos proporciones:

En este caso se contrasta la igualdad de proporciones de una misma clase en dos textos distintos. Se trata de decidir si las diferencias que aparecen en dos textos considerados como muestras son debidas al azar o, en caso contrario, son debidas a una diferencia de estilo. La hipótesis nula es de la forma $P_1 = P_2$, frente a la alternativa $P_1 <> P_2$, donde P_1 y P_2 son las proporciones de una clase cualquiera en cada texto. El estadístico de contraste utilizado en este caso corresponde a la fórmula nº 2.

Como en el caso anterior la hipótesis nula se acepta si el valor del estadístico está comprendido entre -1,96 y +1,96, rechazándose en caso contrario y, como en el caso anterior, el nivel de significación (= 0,05), representa la probabilidad de que la hipótesis nula sea rechazada, aún siendo cierta.

Aplicando este análisis a la comparación de adjetivos calificativos (4) BM y AO, en los que las proporciones son 0,0588 y 0,1488, obtenemos un valor del estadístico de contraste $z = 2,7155$, con lo que la hipótesis nula es rechazada.

c) Contrastar si las proporciones observadas en una muestra son compatibles o equivalentes a las proporciones prefijadas de una población:

Sea n el tamaño de la muestra y $p_1, \ldots, p_k$ las porporciones en la población. Sea $n_1, \ldots, n_k$ las frecuencias absolutas de las clases observadas en la muestra, y $e_1 = n.p_1, \ldots, e_k = n.p_k$ las frecuencias esperadas. Con estos datos planteamos la tabla de contingencia:

Emilia Alonso et al (eds.), La lingüística francesa: gramática, historia, epistemología, Tomo I, Sevilla, 1996

clases	1	2	...	k
frecuencias observadas	n_1	n_2	...	n_k
frecuencias esperadas	e_1	e_2	...	n_k

Ahora planteamos como hipótesis nula que la distribución de frecuencias teórica (la de la población) es equivalente a la distribución de frecuencias esperadas y que, por tanto, las desviaciones entre las observadas y las teóricas son debidas al azar. En este caso el estadístico de contraste corresponde a la fórmula nº 3, donde la variable W sigue una distribución $^2_{k-1}$ (Ji-cuadrado con k-1 grado de libertad) y, como en los casos anteriores, alfa = 0,05, aceptándose la hipótesis nula si W < $^2_{;k-1}$, rechazándose en caso contrario.

Puesto que en la «verificación de una proporción con una valor prefijado» hemos aplicado esta metodología al poema AO para la clase de los adjetivos calificativos, rechazándose la hipótesis nula en esta clase aislada, ahora hemos aplicado la metodología de este apartado al mismo poema, pero considerando todas las clases, obteniendo como valor del estadístico de contraste W = 12,61 < 15,51 ($^2_{0.05;\,8}$), por lo que se acepta la hipótesis nula.

d) Contraste de homogeneidad de muestras.

La prueba de Ji-cuadrado se puede utilizar también para contrastar la homogeneidad de varias muestras aplicadas sobre cada clase, es decir, si h muestras respecto de una misma clase provienen de una misma población (presentan el mismo estilo).

Sean h muestras con $n_1,...n_h$ elementos cada una, las cuales tienen $a_1,...a_h$ elementos de una misma clase en cada una de ellas. Sea $p = (a_1+...+a_h)/(n_1+...+n_h)$, es decir, p representa la proporción de elementos de la misma clase en el conjunto de las h muestras., con lo que los valores esperados serían $p.n_1,...,p.n_h$.

La hipótesis nula es que las muestras proceden de las misma población (homogeneidad de muestras) o, lo que es lo mismo, que las diferencias entre los valores observados en las muestras y los esperados son debidos al azar. Para contrastar esta hipótesis utilizamos el estadístico W, representado en la fórmula nº 4, que sigue una distribución Ji-cuadrado con h-1 grado de libertad, siendo, como en casos anteriores alfa = 0,05. Se aceptará la hipótesis nula si W < $^2_{0.05;\,h-1}$.

Aplicado lo anterior al caso de los adjetivos calificativos obtenemos un valor para el estadístico de contraste de W = 7,48 < 7,81 = $^2_{0.05;3}$, por lo que se acepta la hipótesis nula (aunque próximos a la hipótesis de rechazo).

e) Contraste de homogeneidad de muestras con relación a un valor prefijado de una población.

Se trata en este caso de tomar como valor de p el de la población para la clase objeto de estudio en lugar del valor promedio, procediendo como en el caso anterior.

El mismo ejemplo anterior se ha contrastado con el valor p = 0,1021, es decir, se trata de averiguar si las frecuencias de las muestras para los adjetivos

Emilia Alonso et al (eds.), La lingüística francesa: gramática, historia, epistemología, Tomo I, Sevilla, 1996

calificativos son compatibles con el valor 0,1021. Efectuado el cálculo se obtiene W = 7,47 < 7,81, con lo que se demuestra que la proporción de adjetivos calificativos es «constante» e igual a 0,1021, entendiendo por «constante» que las diferencias que aparecen en cada muestra respecto del valor 0,1021 (el de la población, el del conjunto de los cuatro poemas) son debidas a causas aleatorias.

Pensamos que, al aplicar nuestro método a corpus más extensos, los objetivos propuestos al inicio de este trabajo pueden cumplirse.

En primer lugar, el uso de las herramientas y de nuestra aplicación informáticas permiten llevar a cabo los análisis antes mencionados en un razonable espacio de tiempo y, por otra parte, la utilización del ordenador facilita sobremanera la consulta de concordancias y datos estadísticos, paliando uno de los problemas de este tipo de estudios: su publicación en letra impresa. Un disquete HD puede contener nuestro programa de consulta y, por citar un ejemplo, las concordancias, porcentajes y frecuencias de la obra poética de Villon o de la Chanson de Roland; el usuario, si lo cree conveniente, puede sacar estos datos por impresora o almacenarlos en un archivo para su reutilización en un procesador de textos y, en caso de corpus más extensos, sin ningún problema se puede recurrir a los disquetes comprimidos o al CD ROM.

En segundo lugar, la aplicación de hipótesis estadísticas a los resultados obtenidos ofrece, en nuestra opinión, datos de sumo interés para el investigador. Los tres primeros poemas de las FP llevan la fecha de mayo de 1872 y, el cuarto, del mes de junio del mismo año. Han sido considerados como la cima del arte de Rimbaud y, según Vidal Jover, «estamos en presencia de los 'prodigios de sutileza' que señalara Verlaine; pero es imposible determinar su significación remota» (1972: 361). No hemos realizado un análisis exhaustivo, limitándonos a algunas muestras para experimentar nuestro método y, aún así, dentro de la homogeneidad de estilo, hemos obtenido divergencias que podrán interesar al especialista en busca de esta «remota significación». Pensamos que en los estudios de Lengua y Literatura la conjunción de Estadística y Filología puede, sin duda, contribuir con la aportación de documentos de trabajo que permitan abrir una fecunda vía de investigación.

Fórmula nº 1:

$$z = \frac{p_m - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}} \sim N(0,1)$$

Fórmula nº 2:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_m(1-\hat{p}_m)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

$$\hat{p}_m = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Emilia Alonso et al (eds.), La lingüística francesa: gramática, historia, epistemología, Tomo I, Sevilla, 1996

Fórmula nº 3: $$w = \sum_{i=1}^{k} \frac{(n_i - N_i)^2}{N_i} \sim \chi^2_{\alpha, K-1}$$

Fórmula nº 4: $$w = \frac{1}{p(1-p)} \sum_{i=1}^{h} \frac{(n_i^o - n_i^t)^2}{N_i} \sim \chi^2_{\alpha, h-1}$$

BIBLIOGRAFÍA

DEHENNIN E. (1969): *Cántico a Jorge Guillén. Une poésie de la clarté.* Bruxelles: Presses Universitaires de Bruxelles. Travaux de la Faculté de Philosophie et Lettres. Tome XLI.

DEHENNIN E. (1978): «Des mots clés aux configurations stylistiques. (Surtout à propos de *Maremágnum*)» en *Homenaje a Jorge Guillén.* Massachusetts: Wellesley College, Department of Spanish. Madrid: Ínsula.

FINZI A. y otros (¿?): *Concordancias y frecuencias de uso en el léxico poético de Antonio Machado.* Pisa: Universidad.

GARCÍA WIEDEMANN E. J. (1994): *Concordancias y frecuencias en el léxico poético de los «Proverbios y Cantares» de Antonio Machado.* Granada: Universidad.

GUIRAUD P. (1954): *Les caractères statistiques du vocabulaire.* Paris: PUF.

GUIRAUD P. (1959): *Problèmes et méthodes de la statistique linguistique.* Dordrecht: D. Reidel.

GUIRAUD P. (1969): *Essais de stylistique. Problèmes et méthodes.* Paris: Klincksieck.

JAKOBSON R. (1963): *Essais de linguistique générale,* trad. N. Ruwet. Paris: Ed. de Minuit.

QUEMADA B., MENEMENCIOGLU K. (¿?): *Baudelaire, Les Fleurs du Mal. Concordances, Index et Relevés Statistiques.* Paris: Larousse.

RIMBAUD A. (1984): *Poésies complètes.* Préface, commentaire 12 et notes par Daniel Leuwers. Paris: Librairie Générale Française. Col. *Livre de Poche.*

SABIDO RIVERO V. (1992): *Concordancias de la poesía original de Fray Luís de León.* Granada: Universidad.

VIDAL JOVER J. V. (1972): «Cuatro años de una vida», en *Rimbaud. Poesía completa. Edición bilingüe.* Barcelona: Ediciones 29.

Emilia Alonso et al (eds.), La lingüística francesa: gramática, historia, epistemología, Tomo I, Sevilla, 1996