EL ANÁLISIS LEXICOMÉTRICO EN LITERATURA. PROPUESTA METODOLÓGICA Y APLICACIONES: *CITADELLE, UNE SAISON EN ENFER*

Gabriel María Jordà Lliteras,
Carina Pàmies Valls
y **Carlota Vicens Pujol**
Universitat de les Illes Balears

Los primeros trabajos dedicados al estudio de las frecuencias de las palabras en un texto se remontan a la antigüedad griega, cuando los gramáticos llegaron a elaborar las listas de los vocablos que aparecen una sola vez en las obras de Homero (hápax). A lo largo de nuestro siglo, un creciente interés por el análisis cuantitativo de textos ha supuesto un notable incremento del grado de formalización de los estudios sobre el léxico. Se han llegado a establecer reglas y leyes empíricas formuladas matemáticamente, que pretenden expresar ciertas características inmutables del vocabulario: Guiraud (1950, 1960), Müller (1973), Lafon (1980), Salem (1982), Lebart (1982), Lagarde (1983), etc.

La mayor tradición en la aplicación de los métodos estadísticos al estudio de textos se encuentra en el ámbito francés, en el laboratorio de lexicología de la Escuela Normal Superior de Saint Cloud. Especial mención merecen los trabajos de Guiraud consagrados al análisis de clásicos franceses, fundamentados en la creencia de que la frecuencia es un atributo positivo y concreto de la palabra y que forma parte de su definición, o las aportaciones de Benzécri (1973, 1979, 1981, 1982) dedicadas al análisis de grandes listas de tablas de datos dentro de los análisis de correspondencias.

En sus *Essais de stylistique* P. Guiraud (1969) parte de la idea de que toda obra literaria debe ser entendida como objeto lingüístico o, dicho de otro modo, de que la lengua es la substancia de la obra literaria. Descodificar la lengua de una obra determinada a partir del texto, mostrar que su esencia radica menos en las ideas que esta vehicula que en los hechos lingüísticos que las hacen manifiestas, y todo ello de la manera más objetiva y sistemática posible, tal es el fin de la estilística de Guiraud. No es de extrañar pues que otorgue capital importancia a la estadística, entendida "non seulement comme l'identification et la comparaison des écarts" sino también " de leur interprétation stylistique" porque "à partir du moment où le linguiste conçoit la langue d'une oeuvre comme un code particulier, il y voit non un simple inventaire de formes mais un système de valeurs, dans lequel les signes fonctionnent en opposition et tirent leur sens de leurs relations reciproques au sein de l'ensemble" (1969: 16).

Al servicio de la estilística, la estadística permitirá localizar fácilmente las desviaciones del lenguaje respecto a la norma, entendiendo que dichas desviaciones conforman un estilo, y someterlas al análisis matemático pues, como señala Etxeberría: "la posición de un sujeto emisor ante una realidad sobre la que se pronuncia le lleva a construir su discurso a partir de determinados campos semánticos que definen su actitud. El

María Luz Casal Silva et al. (eds.), La lingüística francesa en España camino del siglo XXI, 2000

vocabulario empleado puede ser tomado como un índice revelador del modo en que los sujetos conciben un determinado hecho, fenómeno o realidad, y el estudio de las frecuencias de este vocabulario, de las frecuencias relativas dentro de unos textos y otros, y las estructuras relacionales que pueden llegar a establecerse entre vocablos y determinadas características de los emisores, nos permiten aproximarnos al contenido de la información textual y llegar a interpetarla" (1995: 140).

Dado que sólo disponemos de las palabras como ámbito de encuentro entre el lector y el autor, pretendemos en esta comunicación presentar una propuesta metodológica que, aprovechando las posibilidades del análisis estadístico y de la informática, nos permita conocer las palabras más significativas del universo léxico de un autor, bien por su homogeneidad a lo largo de toda su obra, bien por su sobre o infrautilización en un momento determinado, e ilustrar nuestra propuesta con el estudio de los nombres-clave de dos obras de la literatura francesa reputadas por su hermetismo: *Citadelle* de Antoine de Saint-Exupéry y *Une Saison en enfer* de Arthur Rimbaud.

Denominamos análisis estadístico de un texto aquellos procedimientos que, mediante el cómputo de las ocurrencias de una o varias unidades verbales básicas permiten realizar, a partir de los resultados obtenidos, algún tipo de cálculo, "de reorganizaciones formales de una secuencia textual y análisis estadísticos con el vocabulario resultante de una segmentación" (Etxebarría, 1995: 145). Dichos análisis se conocen generalmente con el nombre de *lexicometría*. La *segmentación* nos permite el cómputo de las unidades elementales (formas gráficas o lemas) y la reorganización formal de dichas unidades elementales dan lugar al *documento lexicométrico*. Todo este proceso es posible gracias a las herramientas informáticas; en nuestro estudio utilizamos el FRECON (frecuencias y contextos), desarrollado por nuestro grupo de trabajo en la UIB, y el SPAD-T de CISIA (Système portable pour l'analyse des données textuelles / Centre International de Statistique et Informatique Appliquées). En cuanto al análisis estadístico nos servimos del cálculo hipergeométrico desarrollado por Lafon (1980) y de la metodología Chi-cuadrado.

El cálculo hipergeométrico desarrollado por Lafon y calculado con el paquete de programas SPAD-T, permite determinar qué elementos son característicos de un texto dado al realizar un estudio comparativo de varios de ellos, es decir, saber qué palabras aparecen de manera reiterada o son, por el contrario, infrautilizadas. Hablamos de *especificidad positiva* para aquellas formas utilizadas más de lo previsible si se distribuyeran aleatoriamente en todo el corpus, y de *especificidad negativa* para las que se emplean muy poco en relación con su presencia en el corpus global. El problema, como bien señala Etxeberría (1995: 166), consiste "en comparar la frecuencia de los vocables en un texto y la frecuencia de las mismas unidades en un corpus general tomado como referencia... Ha habido distintas aproximaciones estadísticas al estudio de las especificidades, que se han basado en distribuciones teóricas tales como la de chi-cuadrado, la ley normal o la ley de Poisson. Sin embargo es la ley hipergeométrica la que se adapta con exactitud a la población discreta de ocurrencias del vocabulario. Apoyándose en el modelo hipergeométrico, Lafon (1980) ha desarrollado el método de cálculo de las especificidades mayoritariamente empleado en el campo de la lexicometría."

La metodología del Chi-cuadrado es una técnica estadística que se aplica al análisis de frecuencias absolutas y relativas a fin de calcular la homogeneidad de una unidad o conjunto de unidades de uno o varios textos. Se habla de *hipótesis nula*, H_0: $P_1 = P_2$, cuando esta utilización responde a un cálculo aleatorio. Se establece matemáticamente la región crítica (RC), es decir, la zona en la que los cálculos, al ser la estadística la ciencia de lo probable, pueden ser considerados como hipótesis nulas de acuerdo con un *índice de significancia*, y se rechaza la *hipótesis nula* en caso contrario, es decir, cuando el empleo de la unidad o del conjunto de unidades no responde a una utilización aleatoria. La Filología deberá analizar, interpretar y, en su caso, obtener conclusiones de este dato. El *índice de significancia* nos muestra la posibilidad de que se rechace la *hipótesis nula* aún siendo ésta cierta (según el estadístico, según la hipótesis desarrollada); en otras palabras, aunque no del todo correctas desde el punto de vista matemático, el *índice de significancia* nos proporciona el índice de error: una significancia de 0,05 ($\alpha = 0,05$) quiere decir que tenemos una probabilidad de acierto de un 95%.

1. ESTUDIO DE 'CITADELLE'

La aplicación de nuestra propuesta metodológica a *Ctadelle* consta de los siguientes pasos:

1. Traslado del texto a soporte informático en formato ASCII.
2. Desambiguación del texto, marcando las palabras en las que podrían aparecer fenómenos de ambigüedad: aucun' (pronombre), les' (pronombre), leur' (adjetivo), même (pronombre), pas' (sustantivo), que' (pronombre), etc. Este proceso se ha realizado en un total de 162 formas.
3. Para llevar a cabo los cálculos estadísticos debemos establecer las partes del corpus que serán comparadas entre sí y relacionadas con el conjunto global. El texto de *Citadelle*, obra póstuma, las más importante según palabras del propio autor, y redactada a lo largo de casi toda su vida de escritor, queda dividido en 22.305 líneas en el archivo ASCII en cinco partes de 4.461 líneas cada una:

   - la primera parte comprende de la línea 1 a la 4.461, es decir, de la p. 15 a la 133 (capítulos I al XXXI).
   - la segunda parte comprende de la línea 4.462 a la 8.922, es decir, de la p. 133 a la 253 (capítulos XXXI al LXXXI).
   - la tercera parte comprende de la línea 8.923 a la 13.383, es decir, de la p. 254 a la 373 (capítulos LXXXI al CXXXV).
   - la cuarta parte comprende de la línea 13.384 a la 17.844, es decir, de la p. 374 a la 495 (capítulos CXXXV al CLXXXVI)
   - la quinta parte comprende de la línea 17.845 a la 22.305, es decir, de la p. 496 a la 617 (capítulos CLXXXVI al CCXIX).

4. Utilizando una de las opciones de la aplicación FRECON, hemos obtenido el léxico general y el inventario de los nombres, en orden lexicográfico y lexicométrico, de cada una de las partes. Los resultados obtenidos han sido los siguientes:

Primera:

Número total de palabras: 39.149; número de palabras diferentes: 5.794; número total de nombres: 7.215; número de nombres diferentes: 1.945; porcentaje sobre el total de palabras: 18'59433.

Segunda:

Número total de palabras: 39.289; número de palabras diferentes: 5.439; número total de nombres: 6.780; número de nombres diferentes: 1.830; porcentaje sobre el total de palabras: 17'25913.

Tercera:

Número total de palabras: 39.401; número de palabras diferentes: 4.974; número total de nombres: 6.502; número de nombres diferentes: 1.587; porcentaje sobre el total de palabras: 16'48372

Cuarta:

Número total de palabras: 39.547; número de palabras diferentes: 5.287; número total de nombres: 6.658; número de nombres diferentes: 1.656; porcentaje sobre el total de palabras: 16'83759.

Quinta:

Número total de palabras: 40.123; número de palabras diferentes: 5.414; número total de nombres: 7.429; número de nombres diferentes: 1.792; porcentaje sobre el total de palabras: 18'51269.

5. Para realizar con el SPAD-T el cálculo de las especificidades apoyándonos en el modelo hipergeométrico, hemos "reescrito" *Citadelle* con la ayuda del FRECON, elaborando un texto ASCII formado por todos los nombres de cada una de las partes, repetidos según su frecuencia. Los 20 nombres más característicos de cada una de las partes por su sobre o infrautilización, seguidos de su valor test (denominación que el SPAD-T da al resultado del cálculo hipergeométrico) son los siguientes:

Primera:
Especificidad positiva:

père 6.809, généraux 6.693, demeure 5.560, moutons 5.408, demeures 5.013, disparate 4.608, or 4.467, sédentaires 4.455, collaboration 4.338, sel 4.311.

Especificidad negativa:

cérémonial -6.065, clous -5.015, poème -4.674, ami -4.539, lignes -4532, planches -4.260, sentinelle -4.041, roi -4.018, heure -3.894, étage -3.751.

Segunda:
Especificidad positiva:

vie 6.020, paysage 5.522, crête 4.573, loisir 4.566, prière 4.129, plan 4.127, création 4.056, ordre 3.999, Dieu 3.992, montagne 3.970.

Especificidad negativa:

choses -7.667, cérémonial -5.820, condition -3.970, sentinelle -3.862, remparts -3.739, blé -3.696, graine -3.647, part -3.639, besoin -3.639, matériaux -3.606.

Tercera:

Especificidad positiva:

sens 7.993, liberté 7.326, sentinelle 7.248, choses 7.134, temple 5.635, empire 5.268, âmes 4.665, réalité 4.523, contrainte 4.470, objets 4.431.

Especificidad negativa:

soir -5.268, instant -4.903, signe -4.182, tour -3.921, condition -3.869, mot -3.817, façon -3.601, morts -3.601, part -3.545, bonheur -3.532.

Cuarta:

Especificidad positiva:

graine 6.473, part 5.907, ElKsour 5.871, champ 5.394, roi 5.132, géant 4.872, qualité 4.762, change 4.669, sucs 4.412, homme 3.918.

Especificidad negativa:

vie -8.756, cause -6.029, sourire -4.989, femmes -3.785, besoin -3.515, demeure —3.343, corps -3225, or -3.225, frère -3.164, pouvoir -3.164.

Quinta:

Especificidad positiva:

perle 7.991, condition: 7.933, frère 5.759, fête 5.609, cérémonial 5.581, Seigneur 5.525, soif 5.345, amour 4.702, bijou 4.677, échecs 4.617.

Especificidad negativa:

sens -9.755, langage -5.846, homme -5.323, liberté -5.249, hommes -4.887, forme —4.583, porte -4.534, Dieu -4.287, armée -4.116, part -3.726.

6. Mediante el FRECON, búsqueda de las palabras cuyo uso es más homogéneo a lo largo de las 617 páginas de *Citadelle*. Para ello hemos aplicado la metodología del Chi-cuadrado sobre los nombres con un umbral de frecuencia 99, es decir, con un número de ocurrencias de 100 o superior. Diremos que son homogéneas aquellas palabras cuya frecuencia de aparición es similar en todas las partes en que hemos dividido la obra. Con un nivel de significancia de 0,05, de los 45 nombres que se utilizan en 100 o más ocasiones a lo largo de la obra, hemos encontrado los siguientes 17 nombres homogéneos: fois 001.71, désert 001.78, vent 002.23, maison 003.39, peuple 004.62, monde 004.88, arbre 005.32, visage 005.50, enfant 006.04, jour 006.29, vérité 006.42, chose 007.14, mots 007.29, travail 007.49, mer 008.01, pierre 009.07, yeux 009.10. Hay que indicar que la homogeneidad es mayor cuanto menor es el resultado del cálculo estadístico.

2. ESTUDIO DE 'UNE SAISON EN ENFER'

En *Une Saison en enfer* hemos aplicado el mismo método al estudio del nombre y del verbo, considerando como subconjuntos cada uno de los nueve poemas de la obra. Los resultados han sido los siguientes (Después de cada título indicamos las siglas utilizadas en las aplicaciones informáticas):

Une Saison en enfer - SE:

Número total de palabras: 292; número de palabras diferentes: 173; número total de nombres: 56; número de nombres diferentes: 58; porcentaje: 19.86; número total de verbos: 35; número de verbos diferentes: 42; porcentaje: 14.38.

Mauvais sang - MS:

Número total de palabras: 1974; número de palabras diferentes: 812; número total de nombres: 475; número de nombres diferentes: 345; porcentaje: 24.06; número total de verbos: 247; número de verbos diferentes: 178; porcentaje: 12.51.

Nuit de l'enfer - NE:

Número total de palabras: 823; número de palabras diferentes: 396; número total de nombres: 193; número de nombres diferentes: 154; porcentaje: 23.45; número total de verbos: 115; número de verbos diferentes: 76; porcentaje: 13.97.

Délirs I Vierge Folle - VF:

Número total de palabras: 1495; número de palabras diferentes: 629; número total de nombres: 240; número de nombres diferentes: 173; porcentaje: 16.05; número total de verbos: 249; número de verbos diferentes: 175; porcentaje: 16.65.

Délirs II Alchimie du Verbe - AV:

Número total de palabras: 1303; número de palabras diferentes: 650; número total de nombres: 369; número de nombres diferentes: 309; porcentaje: 28.32; número total de verbos: 145; número de verbos diferentes: 116; porcentaje: 11.12.

L'Impossible - LI:

Número total de palabras: 715; número de palabras diferentes: 337; número total de nombres: 132; número de nombres diferentes: 97; porcentaje: 18.46; número total de verbos: 103; número de verbos diferentes: 70; porcentaje: 14.40.

L'Éclair - LE:

Número total de palabras: 255; número de palabras diferentes: 165; número total de nombres: 50; número de nombres diferentes: 43; porcentaje: 19.60; número total de verbos: 36; número de verbos diferentes: 29; porcentaje: 14.11.

Matin - MT:

Número total de palabras: 202; número de palabras diferentes: 135; número total de nombres: 51; número de nombres diferentes: 49; porcentaje: 25.24; número total de verbos: 21; número de verbos diferentes: 21; porcentaje: 10.39.

Adieu - AD:

Número total de palabras: 505; número de palabras diferentes: 301; número total de nombres: 121; número de nombres diferentes: 110; porcentaje: 23.96; número total de verbos: 56; número de verbos diferentes: 47; porcentaje: 11.08.

En cuanto a las formas específicas, por razones de tiempo y espacio nos limitamos a indicar sólo las tres más sobreutilizadas en cada uno de los poemas, seguidas de su valor test:

SE: clef 3.049; festin 3.049; bond 1.823 // appelé 2.933; faire 2.158; joué 1.734.

MS: danse 4.266; race 2.807; amour 2.426 // es 2.696; dit 1.670; trompez 1.558.

NE: enfer 4.313; soif 2.252; horreur 2.252 //entrevu 9.161; veut 2.238; relève 2.233.

VF: époux 3.355; rues 2.767; amies 2.767 //devenir 2.696; voulu 2.184; suis 1.658.

AV: saisons 3.225; châteaux 2.304; bonheur 2.153 // vienne 4.320; éprenne 2.762; boire 2.316.

LI: sagesse 5.311; esprit 4.259; Orient 3.947 // sommes 2.774; habiter 2.318; songeais 2.318.

LE: travail 3.319; ans 3.138; science 1.930 // serait 1.805; éclaire 1.805; tombent 1.805.

MT: Rois 1.880; superstition 1.880; argent 1.880 // désespèrent 2.039; poussent 2.039; ouvrit 2.039.

AD: automne 2.579; souvenirs 2.579; amie 2.579 // éteindre 1.596; fume 1.596; agite 1.596.

Por lo que respecta a la homogeneidad, al no encontrar ningún nombre ni verbo que aparezca en los nueve poemas, hemos estudiado la homogeneidad en el empleo de nombres y verbos en cada uno de ellos y, una vez realizadas las posibles combinaciones, descubrimos una estructura triangular en la composición de *Une Saison en enfer*:

NOMBRES

SE	MS	NE	(SÍ) 5.04	LE	MT	AD
		VF	(SÍ) 2.00	LI		
			AV			

VERBOS

SE	MS	NE	(SÍ)	LE	MT	AD
			4.59			
		VF	(SÍ)	LI		
			1.83			
			AV			

3. CONCLUSIÓN

En nuestra opinión, una vez obtenidas las palabras más representativas de la lengua de un autor a lo largo de su obra, bien por su homogeneidad dentro del corpus global, bien por su sobre o infrautilización en determinadas partes del mismo, cualquier comentario excede el campo de la lexicometría para adentrarse en el de la literatura, lo cual no es objeto ni de esta comunicación ni de este IV Congreso de Lingüística francesa. Ahora bien, para finalizar quisiéramos añadir que, como ciencia de lo probable, la Estadística es para nosotros un instrumento que nos permite obtener unos datos cuantitativos de la realidad objetiva de la lengua para, en etapas posteriores y mediante el análisis cualitativo de estos datos, estudiar los campos estilísticos y la estructura léxica de la obra y poder, así, adentrarnos en la realidad superobjetiva del universo simbólico del autor.

BIBLIOGRAFÍA

BENZÉCRI, J. P. (1973): *L'Analyse des données*. Paris: Dunod.

– (1979): "Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire", in *Les Cahiers de l'analyse des données*.

BENZÉCRI, J. P. et col. (1981): *Pratique de l'analyse des données. Linguistique & Lexicologie*. Paris: Dunod.

BENZÉCRI, J. P. (1982a): *L'Analyse des données 1. La Taxinomie*. Paris: Dunod.

– (1982b): *L'Analyse des données 2. L'Analyse des Correspondances*. Paris: Dunod.

Centre d'Étude du Vocabulaire Français de la Faculté des Lettres de Besançon, avec la collaboration de K. MENEMCIOGLU (s/d): *Baudelaire, les Fleurs du Mal. Concordances, Index et Relevés Statistiques*. Paris: Larousse.

ETXEBERRÍA, J.; GARCÍA E.; GIL J.,y RODRÍGUEZ G. (1995): *Análisis de datos y textos*. Madrid: RA-MA Ediciones.

FRECON *(frecuencias y contextos) programa informático desarrollado para el análisis estadístico-computacional de textos*. Número de inscripción: 00 / 1999 / 3409 Sección: 7 N. Edición: Divulgada: N Clase de obra: Programa de ordenador. Número de R.P.I.....: PM-2233.

GREEN, J. (1990): *Oeuvres Complètes*. Bibliothèque de la Pléiade. VI. Paris: Gallimard.

GUIRAUD, P. (1950): *Les caractères statistiques du vocabulaire*. Paris: PUF.
– (1960): *Problèmes et méthodes de la statistique linguistique*. Paris: PUF.
– (1969): *Essais de stylistique*. Paris: Klinsieck.
LAFON, P. (1980): "Sur la variabilité des fréquences des formes dans un corpus", in *Mots, I.*.
LAGARDE, J. (1983): *Initiation à l'analyse des données*. Paris: Dunod.
LEBART, L. (1982): "L'analyse statistique des réponses libres dans les enquêtes socioéconomiques", in *Consommation I.*.
MÜLLE,R Ch. (1973): *Initiation aux méthodes de la statistique linguistique*. Paris: Hachette.
RIMBAUD, A. (1960): *Oeuvres de Rimbaud*. éd. Suzanne Bernard. Paris: Garnier.
SAINT-EXUPÉRY, A. (1971): *Citadelle*. Le livre de Poche, n° 1.532, 1.533, 1.534. Paris: Gallimard.
SPAD-T, *Système portable pour l'analyse des données textuelles*. Paris: CISIA Éd. 1988.

María Luz Casal Silva et al. (eds.), La lingüística francesa en España camino del siglo XXI, 2000