

HACIA EL ANALISIS AUTOMATIZADO DEL LEXICO DE GONZALO DE BERCEO

Por Manuel Alvar Ezquerria

Cuando hace más de un año preparaba mi comunicación para la I Semana de Estudios Berceanos (1), pude comprobar la escasez de estudios tanto sobre el léxico de nuestro primer poeta conocido como sobre todos los textos medievales, y así lo dije. En estos últimos meses la situación no ha cambiado para nada. Desde la obra de Rufino Lanchetas (2) no existe ningún estudio pormenorizado sobre el léxico del poeta riojano ni de ninguna otra obra del Mester de los clérigos, excepción hecha del *Libro de Apolonio* (3). Tener que tomar la obra de Lanchetas como punto de referencia es algo anacrónico, porque ese trabajo es tan viejo como el siglo, y las cosas han cambiado mucho desde entonces en la manera de concebir los estudios sobre la lengua, por más que en 1900 se hubiera podido exigir un análisis mejor que ése. Cuanto más hoy que los hechos han de ser mirados desde otra perspectiva.

Es cierto, también, que poseemos obras en las que se analiza de una manera más o menos pormenorizada (4) el vocabulario de Gonzalo de Berceo, pero todas ellas no nos dan más que una visión parcial y reducida de ese léxico.

Así fue surgiendo la idea de hacer un análisis profundo de las voces utilizadas por el poeta de Berceo, y de hacerlo lo más rápidamente

(1) "Algunos rasgos léxicos de Berceo y su cotejo con otros poemas hagiográficos", en prensa en el *Homenaje a E. Alarcos Llorach*, Oviedo.

(2) Rufino LANCHETAS: *Gramática y vocabulario de las obras de Gonzalo de Berceo*. Madrid, 1900 (impreso en 1903).

(3) Ed. de Manuel Alvar, 3 vol., Madrid, 1976. Aquí nos interesa especialmente el t. III.

(4) Resaltaba en mi trabajo de la nota (1) los Indices de Gerhard KÖRBERSTEIN a la *Estoria de San Millán* (Münster, 1964), y de Pompilio TESAURO al *Martirio de San Lorenzo* (Liguori-Nápoles, 1971), más completo.

posible. No voy a referirme ahora a otros trabajos parecidos que se han realizado mediante la ayuda de los ordenadores electrónicos, pero sí es verdad que deseamos saber los términos utilizados por Berceo exhaustivamente y pronto, será imprescindible la automatización del trabajo en la mayor parte de sus fases. Además, el estado actual de nuestros conocimientos no puede dejar esa enorme tarea como una finalidad que concluya en sí misma, sino que se le debe buscar una proyección mucho más amplia, ya sea en la inclusión de ese análisis en el futuro Diccionario del español medieval, o en el más ambicioso Tesoro de la lengua española, sin olvidar otros aspectos, tales como el estudio de las estructuras gramaticales de tal o cual obra, o del conjunto de las obras, etc. Por todo ello hay que cuidar con gran detalle la preparación de los materiales para su posterior estudio y aprovechamiento.

El primer aspecto a tener en cuenta antes de programar el análisis es el de los textos que se van a manejar, tema de singular importancia para que la obra no quede invalidada en unos pocos años, conforme se puedan mejorar las ediciones. Es consenso común, con escasas discrepancias, que las mejores ediciones de los textos de Gonzalo de Berceo son las realizadas por Brian Dutton (5). En estos momentos poseemos ya siete títulos preparados por dicho investigador en tres volúmenes, materiales suficientes para trabajar durante algún tiempo. Naturalmente, si la rapidez de nuestras investigaciones fuera superior a la impresión de todas las obras de Berceo en las ediciones de Dutton, escogeríamos entre las publicadas la que reuniera mejores condiciones.

Es conveniente, en nuestro análisis, dar cuenta no sólo de las voces recogidas en la edición crítica utilizada, sino también de las variantes de los distintos manuscritos, para lo cual son de una inestimable ayuda las ediciones del investigador inglés. Lo que de ningún modo será posible es conceder a esas variantes la misma importancia que a las formas empleadas en la edición crítica, porque el autor escribió (si la edición crítica ha alcanzado sus propósitos) una de esas formas y no las demás, por más que éstas tampoco puedan ser olvidadas, ya que existen. La solución para los análisis estadísticos radica en hacerlos sólo a partir de la edición crítica, de manera que

(5) Todas ellas publicadas en Londres: la *Vida de San Millán de la Cogolla* (1967), los *Milagros de Nuestra Señora* (1971), y el *Duelo de la Virgen*, los *Himnos*, los *Loores de Nuestra Señora* y los *Signos del Juicio Final* (1975).

los resultados definitivos no queden falseados. Ello no impide dar cuenta del léxico de las variantes, indicando su lugar y procedencia, esto es, el número de estrofa y verso y el manuscrito.

En principio, los análisis serían los habituales en este tipo de trabajos:

- a) índice alfabético de frecuencias absolutas
- b) índice alfabético de frecuencias relativas.
- c) índice numérico de frecuencias absolutas (las relativas ya no son necesarias y se ahorra espacio).

Estos tres índices se efectúan sobre palabras-texto.

- d) concordancias lematizadas
- e) índice numérico de frecuencias absolutas por lemas
- f) índice numérico de frecuencias relativas por lemas

Los índices alfabéticos de frecuencias podrían ir fundidos en las concordancias (punto *d*).

- g) índice de rimas.

Si queremos conocer el léxico de Gonzalo de Berceo de una manera global, habrá que reunir todos esos resultados parciales de cada obra para dar cuenta de la totalidad de las voces registradas. Entonces bastaría con un número menor de índices. El primero de ellos consistiría en una lista de todos los lemas de la obra de Berceo, con las indicaciones de sus frecuencias tanto absolutas como relativas. En el segundo incluiremos todas las formas bajo las que se puede presentar un lema, también con la notación de sus frecuencias. El tercero será una lista de todas las rimas encontradas en la obra del poeta riojano.

Los dos primeros índices de los que acabamos de hablar podrían quedar fundidos en uno solo, incluyendo las variantes formales de cada lema bajo su entrada; además, de este modo, se presentarán las frecuencias absolutas y relativas con respecto a la totalidad de la obra de Berceo, pero también la relativa en comparación, sólo, con las otras formas del mismo lema, información posiblemente de menor interés, pero necesaria para el estudio del léxico. En estos índices habría una entrada para cada una de las formas, con una indicación que remitiera al lugar donde se analiza la forma en cuestión. Fuera la que fuera la solución adoptada, tras cada forma habrán de reseñarse todas y cada una de las localizaciones exactas en los escritos de

nuestro poeta, pues suprimir alguna de las formas o de sus localizaciones sería presentar unos resultados falsos. Para que ese apartado no sea largo en exceso se buscará un sistema de abreviaturas para dar la referencia de la obra, estrofa y verso.

Por otro lado, dentro de la producción de Gonzalo de Berceo, existen obras que tienen un marcado carácter fragmentario, tales como los *Milagros*, o los *Himnos*, y que son merecedoras de un análisis más pormenorizado, por varios motivos bien evidentes y de un interés que no voy a enumerar. Mantendremos para estas composiciones los principios expuestos hasta aquí, pero además se establecerán otros índices para cada una de las partes en cuestión. A saber:

- a) índice alfabético por palabras-texto de frecuencias absolutas
- b) índice alfabético por palabras-texto de frecuencia relativas al conjunto de la obra
- c) índice alfabético de palabras-texto de frecuencias relativas a la parte tratada
- d) índice numérico de frecuencias absolutas, por palabras-texto
- e) índice alfabético de palabras-texto de frecuencias relativas a
- f) índice alfabético de frecuencias relativas al conjunto de la obra, por lemas
- g) índice alfabético de frecuencias relativas a la parte tratada, por lemas
- h) índice numérico de frecuencias absolutas, por lemas.

En los índices señalados desde *e* hasta *h* sería conveniente indicar, tras cada lema, las diferentes formas halladas, con las mismas indicaciones estadísticas que se ofrecerán por los lemas. Quizás esto impidiera ver con claridad el resto de las indicaciones, por lo que las variantes formales pueden ser reducidas a los índices de frecuencias absolutas.

De la manera expuesta se daría cuenta de un modo fiel del léxico utilizado por Gonzalo de Berceo. La obra sería de una enorme utilidad, en mi opinión, porque podríamos ver simultáneamente los resultados del conjunto de los escritos del poeta de la Virgen, y los análisis parciales de cada obra, e incluso de cada una de las partes en aquellos textos que lo permitieran.

El método a utilizar, como ha quedado dicho antes, tendría una ayuda eficaz y primordial en los ordenadores electrónicos, sin olvidar

que algunos pasos de nuestro análisis habrán de darse manualmente, pues las circunstancias por las que atraviesa la investigación de nuestro país —en especial la economía— no permiten, por el momento, alardes en la técnica del análisis, por más que sea posible hacerlos y que se hayan efectuado con éxito en otros países. De todas formas, la automatización total en este tipo de trabajos es aún una meta lejana, si es que alguna vez se puede lograr alcanzarla.

En otros trabajos míos he expuesto la manera de llevar a cabo estudios semejantes al que propongo ahora (6), por lo que me voy a ahorrar el repetirlo de nuevo. Sí quisiera decir que los principios son buenos a la vista de los resultados obtenidos.

El proyecto del análisis automático del léxico de Gonzalo de Berceo se ha puesto en marcha de una manera muy modesta en el Centro de Proceso de Datos de la Universidad de Málaga, en colaboración con el Departamento de Lingüística Sincrónica de la misma Universidad. Digo de una manera modesta porque los presupuestos con que contamos no nos permiten hacer alardes de ningún tipo, y porque el ordenador de que disponemos (un IBM 5100) tiene una capacidad muy reducida, por lo que trabaja en teleproceso con una unidad central mucho más potente situada en Madrid. Es posible que nuestro Centro de Proceso de Datos cuente en los próximos meses con un terminal del gigantesco Univac que posee el Ministerio de Educación y Ciencia, y que entonces se agilicen todos los trabajos que tenemos proyectados en estos momentos, y que coparían la capacidad del pequeño ordenador existente hoy.

La primera obra que queremos tratar son los *Milagros de Nuestra Señora*, con mucho el escrito más conocido y estudiado de Berceo, motivos por los cuales el análisis de su vocabulario no se puede demorar más. Por otro lado, ofrece la posibilidad de ir trabajando poco a poco, milagro por milagro, de modo que los problemas que surjan puedan ser resueltos en pequeña escala, con lo que al fundir los análisis parciales no resultarán nuevos inconvenientes, y tendremos el camino expedito al enfrentarnos con los otros escritos berceanos, ya sean de gran extensión ya de un pequeño número de ver-

(6) Véase mi *Proyecto de lexicografía española*, Barcelona, 1976, en especial el capítulo "La asistencia de los ordenadores en la elaboración de obras lexicográficas" (pág. 203 y sigts. del libro). Cfr. asimismo la comunicación que presenté a la reunión sobre lingüística e informática (Madrid, 16-17 de noviembre de 1977), titulada "Trabajos realizados sobre léxico medieval en el ECUC":

sos. Y si por cualquier motivo fuera necesario interrumpir la ejecución de la obra poseeríamos los resultados definitivos del estudio de cada uno de los milagros.

En estos momentos se ha preparado la preedición de la “Introducción” y del primer milagro, que ofrecen pocas dificultades, salvo las habituales en este tipo de trabajos. Los signos que se van a utilizar son:

- (^) tras una vocal indica que esa vocal es portadora de la tilde acentual. Tras una *n*, si se encuentra algún caso, indicará que debe ser leída como ñ.
- (..) es el signo de diéresis, que tiene que ir postpuesto a la vocal, pues es imposible sobreponerlo.
- (◊) sustituirá a la *ç*.
- (/) señalará el final de verso.
- (°) aparecerá en lugar del punto volado utilizado por Dutton en *diol*, por ejemplo.

Las mayúsculas irán subrayadas, menos cuando sean mayúsculas por posición, que no llevarán indicación especial alguna, pues si se conocen las reglas ortográficas será fácil deducir cuándo una letra es mayúscula y cuándo no lo es; de esta manera no se entorpecerán otros posibles trabajos posteriores, como puede ser el estudio de los nombres propios.

Hay otros signos que serán utilizados en el tratamiento, pero que no aparecerán en los listados definitivos, como puede ser (††), utilizado para indicar el final de estrofa y de contexto, ya que para facilitar el trabajo se considerará como contexto de cada palabra toda la estrofa donde se documenta. Mediante el signo de igual (=) —que tampoco será escrito— se unen los elementos de las lexías complejas estables de la terminología de B. Pottier. Se consideran dentro de este grupo de lexías los complejos:

san Pedro (17 *d*)
sant Illefonso (57 *a*, 66 *b*)
Virgo María (30 *c*, 31 *d*)
Ave María (53 *b*)
Dios Nuestro Sennor (68 *d*)

así como el sintagma latino “post partum et in partu” (20 *c*), o el otro “requiescat in pace cum divina clemencia” (99 *e*), donde *cle-*

mencia, por su grafía, puede despistar, y el refrán “nin Sancho nin Domingo, nin Sancha nin Domenga” (38 d). Más dificultades ofrece el análisis de *fuesse ido* en el verso 70 a, pues aparece “fuesse Siagrio tan adelante ido”; sin embargo, la lematización manual permite el estudio por separado del verbo y de los otros tres elementos (7).

Aún no han aparecido otros signos para los que habrá que buscar soluciones en consonancia con las posibilidades de la máquina, como pueden ser los de interrogación y exclamación. Para los de interrogación será preciso abrir y cerrar con el mismo signo (?), y para los de exclamación utilizaremos la barra vertical (/), siempre que no se nos ofrezcan mejores perspectivas por parte de las máquinas.

Los lemas aparecerán bajo la forma habitual en los trabajos lexicográficos (verbos en infinitivo, adjetivos en masculino singular, etc.). Los artículos y pronombres personales tendrán separados los géneros, como hacen muchos diccionarios, para poderlos comparar entre sí. La grafía de los lemas será la corriente en nuestros días, a fin de facilitar al posible usuario la búsqueda de los términos. Ello no impide que exista una entrada para cada variante gráfica que enviará al lema correspondiente donde esté recogida esa variante. Si la voz estudiada no ha llegado hasta ahora, no se modernizará la grafía del lema; y si existen varias grafías se elegirá la más próxima a las actuales, tal es el caso de las alternancias de las vocales, por ejemplo. En el momento en que haya dudas para la reconstrucción de los términos se dejarán bajo la forma que tienen en el texto, como puede suceder con *maestr'* (¿se lematiza *maestre* o *maestro*?).

Los textos están siendo introducidos en la memoria de la máquina mediante “cassettes”, único soporte que puede leer en nuestro Centro de Proceso de Datos el ordenador IBM 5100. El método de los cassettes ofrece algunas ventajas sobre las fichas, utilizado habitualmente en este tipo de trabajos: el poder ir leyendo en la pantallita lo que se escribe en el mismo momento de hacerlo, de manera que la corrección de errores puede ser instantánea; otra ventaja es el no necesitar repetir al comienzo de cada verso las siglas que identifican el texto, etc. En nuestro caso, la identificación la hacemos mediante las siguientes siglas:

(7) Más información en los trabajos citados en la nota anterior.

GB para el autor, Gonzalo de Berceo

MI para el texto, los *Milagros*

IN para la "Introducción", y en su lugar, cuando se analicen los milagros, de 01 a 25, con números arábigos, pues hacerlo con romanos sería complicar excesivamente el trabajo.

Por ahora sólo queda la obtención de los primeros resultados parciales para conducir la investigación por buen camino. La obra parece realizable, siempre que se cuente con el apoyo económico necesario.