LA APLICACIÓN DEL ORDENADOR A LA REALIZACIÓN DE CONCORDANCIAS

Elisa M.ª Ferreira Priegue

El proceso electrónico de datos, auxiliar inestimable en trabajos que requieren el tratamiento y análisis de grandes volúmenes de información, no supone una novedad, a estas alturas, en el campo de la filología. Existen en Europa instituciones especializadas en el procesamiento electrónico de materia lingüística, como el L.A.S.L.A. (Laboratoire d'Analyse Statistique des Langues Anciennes de l'Université de Liège) y el Centre d'Analyse et Documentation Patristique de l'Université de Strasbourg, por no citar más que dos. En otros casos, se ha recurrido a los servicios de otros centros de cálculo, con los inconvenientes que esto supone: adaptación a lenguajes no específicos de estas aplicaciones, limitaciones de los signos de impresión y desconocimiento de la materia por parte de programadores y operadores y, recíprocamente, de las especificaciones y exigencias del proceso de datos por parte de los estudiosos.

Las aplicaciones del ordenador a la lingüística conocieron un gran «boom» en la década de los 60, hasta que se llegaron a conocer mejor sus posibilidades y limitaciones. Hoy en día, los métodos que entonces se habían desarrollado con objetivos mucho más ambiciosos (traducción automática, análisis gramatical y semántico) se utilizan en otras aplicaciones muy típicas de este campo, como la confección de índices, léxicos y concordancias. Fuera del campo propiamente humanístico, pero muy vinculado a la lingüística, las técnicas concebidas con miras al análisis semántico han sido muy útiles para el perfeccionamiento de ese instrumento informático que es el «thesaurus».

Uno de los trabajos más sencillos y rentables dentro de estas aplicaciones actuales son los índices de frecuencia del léxico empleado por un autor. Estos índices, sometidos a diversos pasos de clasificación con criterios cronológicos, lingüísticos, geográficos, sociológicos y biográficos —por no citar más que unas cuantas posibilidades—, pueden arrojar mucha luz sobre la personalidad y estilo del autor y el estado de la lengua en su época, haciendo en un par de horas de máquina un trabajo rutinario que significaría muchos meses de trabajo humano. La complejidad de estas estadísticas abarca, por lo tanto, una escala muy amplia, desde un simple índice alfabético, sin más pretensiones, como el *Index Nepotianus* de Leiniecks[1], realizado al nivel básico de un listado alfabético del léxico de Nepote, con sus correspondientes referencias, hasta trabajos más sofisticados como la *Biblia Patristica*[2], que pone al descubierto la utilización de los textos bíblicos en la obra de los Padres de la Iglesia, o los índices lematizados de las comedias de Plauto[3], con índices de frecuencia de palabras, directo e inverso, de toda la obra y por personajes, con varias clasificaciones por categorías gramaticales, declinaciones, verbos en todas sus formas, etc.

Las concordancias suponen un nivel más avanzado en este proceso de análisis textual, y plantean otros problemas más graves, que rebasan los alcances de un simple índice.

El primero de ellos es el de la selección del contexto. En los sistemas mecánicos de indización, son dos los principales tipos de índices que recogen texto en torno a la palabra clave, y lo presentan de dos maneras:

El KWIC (Keyword-in-Context) posiciona la palabra clave en el centro de la línea de impresión, recogiendo un determinado número de posiciones de impresión a ambos lados de la misma. Es un sistema de búsqueda cómodo y rápido por su gran impacto visual, pero la recogida de contexto es demasiado mecánica. Éste es el principal problema de estos índices para un contexto literario, aun cuando son muy eficaces en textos breves (por ejemplo, índices de títulos de libros y artículos, en que las palabras clave determinan su materia). En algunos casos, este problema se puede salvar con bastante éxito, recurriendo a medios extraordinarios y que estén dentro de las posibilidades del ordenador concreto que se esté utilizando, como se hizo en la concordancia de Livio de David

[1] LEINIEKS, Valdis, *Index Nepotianus*, Lincoln, University of Nebraska, 1976.

[2] *Biblia Patristica. Index des citations et allusions bibliques dans la littérature patristique*, Paris, C.N.R.S., 1975.

[3] MANIET / PAQUOT, *Plaute. Amphitryon. Index verborum. Lexiques inverses. Relevés léxicaux et grammaticaux*. Hildesheim-New York, Georg Olms, 1970.

Packard[4], conectando al ordenador una unidad de fotocomposición que permitió, aparte de una bella presentación tipográfica, la recogida de un contexto casi cuatro veces más amplio que si se hubiera impreso directamente en ordenador, y que en esas condiciones es bastante operativo.

El KWOC (Keyword-off-Context) es más sensible a la hora de recoger el contexto y, dentro de las consabidas limitaciones de la línea de impresión, procura que tenga el mayor sentido posible. Pero como esto, naturalmente, impide que la palabra clave quede centrada, se la entresaca al margen, a modo de lema, y en el lugar que le correspondería dentro del texto se deja un espacio en blanco o algún otro signo que ahorre posiciones de impresión. Lo fatigoso y antiestético que puede llegar a ser este índice para su consulta está ejemplificado en las *Concordantiae Senecanae* de Busa y Zampolli[5], que evidencia además el empleo de un ordenador no preparado para estos menesteres. El contexto es reducidísimo y, por falta de signos adecuados en la impresora, la notación es muy complicada y se han tenido que emplear caracteres extraños en lugar de algunos signos ortográficos (por ejemplo, ? se representa ≠).

En ambos sistemas de indización, el problema del contexto sigue sin una solución satisfactoria. El texto, por amplio que sea, siempre queda mutilado o redundante, y por lo tanto no cumple con su requisito más elemental, que es el de dar pleno sentido a la palabra estudiada.

Otro problema de difícil solución mecánica es el de la lematización. Es el que requiere más trabajo humano a la hora de realizarlo en ordenador, y es demasiado fuerte la tentación de adoptar para su confección un instrumento informático que a primera vista parece el indicado para este proceso: el «thesaurus».

Las limitaciones de este sistema se hacen evidentes en la concordancia de Marcial realizada recientemente por Siedschlag[6], que recoge el contexto utilizando el formato de indización KWIC. En cuanto a presentación tipográfica, el texto aparece tal como salió de la impresora del ordenador, con sus característicos tipos en mayúscula, que le restan dignidad y claridad de lectura. Al querer lematizar la concordancia, se ha sometido el vocabulario de Marcial a un tratamiento que lo reduce a los estrechos límites de un «thesaurus» informático. Nos daremos cuenta de lo

[4] PACKARD, David W., *A Concordance to Livy*, Cambridge, Mass., Harvard University Press, 1968. Siguiendo muy de cerca el procedimiento de Packard, es también un buen trabajo el de MCCARREN, *A Critical Concordance to Catullus*, Leiden, Brill, 1977.

[5] BUSA, R. y ZAMPOLLI, A., *Concordantiae Senecanae*, Hildesheim-New York, Georg Olms, 1975.

[6] SIEDSCHLAG, Edgar, *Martial-Konkordanz*, Hidlesheim-New York, Georg Olms, 1979.

que esto significa si recordamos qué es, en el lenguaje de los documentalistas, un «thesaurus». Explicado de una forma muy básica, consiste en un léxico de palabras clave, en un lenguaje que puede ser natural o controlado, pero que en cualquier caso está reducido a unas convenciones por necesidades funcionales. Este léxico se emplea en un doble juego combinado: listado alfabético y «thesaurus» propiamente dicho, este último con una determinada disposición lógica de las palabras por diversos tipos de asociación: jerarquía, contigüidad, afinidad, etc. Es decir, algo similar a un diccionario ideológico. Se emplea para recuperar información sobre una materia dada en un almacenamiento de datos. Es obvio que, para un análisis de textos literarios, este sistema que, por ejemplo, ignora las partículas y agrupa los adverbios con sus adjetivos correspondientes resulta enormemente insensible, aun contando con un buen aparato de referencias cruzadas, que la concordancia de Siedschlag no tiene. Este problema lo resuelve airosamente el autor, aconsejando que, en caso de no encontrar el lema deseado, se recurra a los léxicos de base por los que se ha guiado —El *Thesaurus Linguae Latinae*, el *Nuevo Diccionario de Oxford* y el *Lewis & Short*—, que tienen este tipo de referencias y pondrán al estudioso sobre la pista de la palabra que busca.

De todas formas, prescindiendo de la insuficiencia de los contextos —mal común a todos estos trabajos— y del hecho de no recoger variantes, la obra hubiera resultado un buen índice, de fácil lectura y recuperación gracias al sistema KWIC, y muy manejable, a no ser por una serie de fallos que presenta, unos emanando del propio método y que se pueden ir paliando gracias a las advertencias del autor en su introducción, y otros, los más graves, consistentes en descuidos y omisiones que hacen de esta concordancia un instrumento de consulta de fiabilidad bastante dudosa.

La concordancia se ha quedado a medio camino entre un índice y un «thesaurus». Para el primero resulta redundante un contexto de muy poca utilidad; del segundo falta la cuidadosa combinación entre lista alfabética y «thesaurus» propiamente dicho (que en este caso correspondería a un juego de índice alfabético y léxico lematizado), todo fuertemente estructurado por un sistema de «links» y referencias cruzadas y de los demás dispositivos de recuperación y notación de que tiene que constar un *corpus* de lenguaje indizado.

Tenemos, por ejemplo, el problema de los sinónimos y homógrafos, caballo de batalla de primera magnitud en un «thesaurus», y para cuya solución hay previstos unos tratamientos complejos y laboriosos. En el caso de los homógrafos, las reglas que da el autor advierten que los ad-

verbios y conjunciones no se distinguen de las formas pronominales homógrafas, y aparecen todos juntos bajo el lema correspondiente al pronombre; que no se hace distinción entre *cum* preposición y conjunción, y que las formas de *quis* o *qui* aparecen todas bajo el lema QVIS. Así, por ejemplo, bajo HIC aparece también *hic* (adv.); bajo QVIS, todas las formas de *quam, quod, quo,* etc. Aparte de esto, se dan numerosas confusiones, como en ACVS, -i y ACVS, -us. En el verbo EO, las formas *eas* e *isti* aparecen en IS, ea, id y en ISTE, -a, -ud.

Por la índole del trabajo, la cuestión de los sinónimos apenas se plantea aquí. Sin embargo, términos asociados pueden dar lugar a confusiones. Como el sistema de lematización ignora los nombres compuestos y lista palabra a palabra, el nombre *Anna Perenna* recoge bajo el lema ANNA, *Anna* y *Perenna; Perenna* es imposible de encontrar para quien lo busque en la P.

Esto ya es de por sí un fallo grave; pero aún peores son los numerosos errores que se encuentran a lo largo de toda la concordancia y de los que se dan aquí algunos botones de muestra:

Los nombres propios van separados de los comunes, distinguiéndose con un asterisco: Sin embargo, figuran entre los nombres comunes:

— AVGVSTVS.
— AETIA, como nombre común en AETION.
— ARCTOVS, -a, -um (a diferencia de otros gentilicios y adjetivos de origen geográfico, que figuran como nombres propios).
— HELICE.
— CORACINVS, bajo un lema que mezcla el nombre propio y el común.
— FLAVVUS (X 104, 1) está bajo FLAVVS, -a, -um.
— FORTUNATUS, n. pr., está con el adjetivo bajo el lema FORTVNO.
— GALLA (n. de mujer) bajo el gentilicio GALLI, junto con GALLIA, GALLICVS, GALLICANVS.
— GALLVS (n. de varón) y GALLVS (sacerdote de Cibeles), aparecen en dos lemas separados, pero con las acepciones mezcladas.

Muchos de estos nombres propios, según los criterios de recuperación de los «thesauri», van lematizados por el adjetivo, nombre común o verbo del que se supone derivan, y que aparece o no en el texto de Marcial en su forma primitiva, lo que no facilita su búsqueda. Así, hay que buscar FORTVNATVS en el verbo FORTVNO; GRATIANA en GRATIVS; bajo el lema HISPANI, encontramos HISPANI, -orum, HISPANVS, -a, -um, HISPANIA, -ae e HISPANIENSIS, -e. Bajo APOLLO están APO-

LLINARIS, APOLLINVS, APOLLINEVS, -a, -um, APOLLO, APOLLODORVS y APOLLODOTVS. Bajo ANTIVS, ANTISTIVS, ANTONIVS y ANTVLLA, etc. Sin embargo, y contra esta convención, separa CERVSSA y CERVSSATVS, HECTOR y HECTOREVS; COCLEA y COCHLEA aparecen como dos lemas distintos, y bajo COCHLEA están COCLEARE y COCLEARIVM.

Se trata de una alfabetización excesivamente mecánica, que no parece haber tenido una adecuada supervisión de los lemas; esto conduce a una excesiva condensación de derivados remitidos a un lema primitivo, a veces inexistente en el texto, como CERO para CERATO y CLVSIVM para CLVSINVS. Pero la regla no siempre es respetada, ya que, por ejemplo, FESSVS no está lematizado, como debiera, por el verbo FATISCOR.

Esta sensación de descuido en la confección de la concordancia llega al máximo cuando nos encontramos con una serie de errores de bulto que parecen remontar al proceso de clasificación, y que consisten en mezclas de palabras pertenecientes a distintos lemas. Los más llamativos son el de la página 184, en la que aparecen intercalados, bajo el lema CLEMENTIA, COLO, parte de COLLIGO, COLOR, parte de COMA, todos fuera de su lugar correspondiente y sin lematizar; en la p. 192, bajo el lema COMPINGO, están entremezclados COMPINGO, parte de COMPLEXVS, CONSCIENTIA y parte de CONSVL, etc. También hay otra mezcla de FORAS, FORIS (adv. y prep.), parte de FORIS, -is y parte de FORVM, -i.

Hay también omisiones de lemas enteros, como por ejemplo CIRCEII, -orum, COLICVLUS, AMILLVS, etc. Las omisiones de citas dentro de lemas y la numeración incorrecta de las referencias se dan en ACCIPIO, AEDES, CONTENTVS, COMA, CITHARA, FATIGO, FELLO, FERA, -ae (que, dicho sea de paso, está bajo un único lema con FERVS, -a, -um), y otros demasiado numerosos para citar. En cambio, CVMQVE está incorporado en dos sitios: bajo el lema CVMQVE e incluido en CVM.

En estas condiciones, la utilidad de esta concordancia como instrumento de trabajo es poco menos que nula, y puede conducir a errores peligrosos. Pongamos por caso que un estudioso buscando los casos de empleo de la palabra *coma* en Marcial, la busque en su lugar correspondiente, en la página 189-90. Se encontrará con 45 entradas; pero hay tres más, de las cuatro en que la palabra aparece en acusativo, que están extraviadas en la página 184, bajo el lema CLEMENTIA, donde, por supuesto, no se le ocurrirá irlas a buscar, y el número de entradas bajo el lema principal le parecerá suficiente. Las conclusiones a que llegue pue-

den quedar totalmente invalidadas. No hay que decir que todos estos trastrueques y omisiones falsean también los datos estadísticos de los índices de frecuencia.

Es evidente que, pese a las deficiencias de la mecanización para este tipo de trabajos, esta concordancia pudo haberse realizado de una forma más cuidada, a la vista de otros precedentes ya citados. En cualquier caso, una concordancia con una lematización concienzuda y flexible, y con un contexto coherente, sigue siendo una empresa en la que ningún ordenador puede suplir el trabajo personal del autor, aunque le resulte un valioso auxiliar para las tareas más rutinarias, así como para las estadísticas adicionales que desee obtener.