La "Informática Humanística": notas volanderas desde el ámbito hispánico

José Manuel Lucía Megías
Universidad Complutense de Madrid

1

Seguir hablando y hablando a principios del siglo XXI sobre la posible influencia de las nuevas tecnologías en el ámbito de las humanidades (y de la filología si queremos ser más específicos), o de cómo influirán las nuevas técnicas y herramientas informáticas en nuestras disciplinas académicas, tanto en su enseñanza como en su investigación, se ha convertido en un lugar común de muchos coloquios y seminarios científicos. Pero este hablar supone un seguir dando la espalda a una realidad: no se trata tanto de evaluar cómo influirá en las humanidades en un futuro, más o menos cercano, más o menos lejano, una tecnología –que ya no podemos seguir considerando *nueva*–, sino de aceptar cómo ésta ya ha modificado aspectos básicos de nuestra vida cotidiana y profesional, y cómo hemos de ofrecer ya nuevos conocimientos y herramientas dentro de las aulas académicas para que el alumno de cualquier disciplina humanística pueda moverse con naturalidad en el nuevo espacio profesional y científico de los próximos años. No podemos seguir discutiendo sobre problemas de nomenclatura universitaria propios del siglo XIX cuando el reto tecnológico está llamando –como sólo a principios del siglo pasado sucedió– a las puertas de nuestras disciplinas.

Incipit, XXIII (2003), 91-114

La *Informática Humanística* (la *Humanities Computing* en el ámbito anglosajón[1]) se presenta como una disciplina científica que, en la actualidad, se encuentra (casi) ausente de los planes de estudio universitarios, a excepción de una de sus ramas, la lingüística computacional, que ha conocido un enorme desarrollo en los últimos tiempos. Pero las fuentes de la *Informática Humanística* no hay que buscarlas en los últimos decenios del siglo XX, sino en los años cuarenta de la centuria, cuando el padre Roberto Busa solicitó a J. Thomas Watson, por aquellos años presidente de IBM, su apoyo para crear una versión electrónica de las obras de Santo Tomás de Aquino, con la finalidad de obtener un completo *index*, más fiable que los índices y concordancias manuales existentes hasta entonces, que ha visto la luz con el título de *Index Tomisticus: sancti Thomae Aquinatis operum omnium indices et concordantiae*[2]. El proyecto, que nace de la imposibilidad humana de encontrar referencias internas en la extensísima obra del santo, ha ido transformándose a medida que la informática ha ido perfeccionando sus herramientas y modos de difusión. Las cifras son más que elocuentes: en un primer momento, se trabajó con tarjetas perforadas, llegando a 12 millones de ellas (que ocupaban en estanterías el equivalente a noventa metros de largo, por 1'20 de ancho y 500 toneladas de peso); después se pasó el material a las cintas magnéticas, llegándose a contabilizar 1600 kilómetros de cinta. El resultado científico se publicó en primer lugar en papel –de 1974 a 1980–, constituyendo 56 tomos, con más de setenta mil páginas impresas; a partir de 1992, se ha pasado todo el material a formato electrónico, a varios CD-ROMs, mientras que en la actualidad, gracias a la mayor densidad de acumulación de los discos ópticos, se ha conseguido reducir a uno solo.

El ejemplo del *Index Tomisticus* nos habla de un precursor y de una historia, pero también de cómo la informática, con su enorme

---

[1] Que ha dado lugar en los últimos años a órganos periódicos de difusión, entre los que destacamos los siguientes, en donde el lector interesado podrá encontrar las pertinentes referencias bibliográficas: *Literary and Linguistic Computing, Computing and Humanities* y *Journal of the Association for History Computing*...

[2] Una historia del proyecto, en boca del propio padre Busa, puede consultarse en Internet, en la entrevista que en 1995 le hizo la RAI: http://www.mediamente.rai.it/home/bibliote/intervis/b/busa.htm.

desarrollo en las últimas décadas, se ha convertido en un instrumento esencial para la creación de herramientas para los estudios humanísticos. Y esta dinámica se ha vuelto revolucionaria a partir de la década de los noventa, con la difusión y extensión de Internet y de la Web, gracias a la que se está dibujando un nuevo marco social, que se denomina "sociedad de la información", en donde la tecnología se va haciendo cada vez más "humana"; cada vez más se va introduciendo en nuestros ámbitos personales y profesionales, por lo que parece inevitable un espacio de relación entre esta *nueva* disciplina –que desde hace tiempo goza de su autonomía dentro de las facultades tecnológicas– y las *antiguas* humanidades. Ese es el espacio que debe ocupar la *Informática Humanística* desde una visión amplia de sus contenidos; una visión amplia, que parte del conocimiento desarrollado en los dos últimos siglos por la ciencia filológica.

2

La *Informática Humanística* se ha quedado, hasta ahora, a las puertas de la Universidad hispánica: las enseñanzas se dispersan entre los heterogéneos programas de doctorado y en cursos de diferente categoría en facultades de Humanidades (como la de la Universidad de Castilla-La Mancha) o dentro de las titulaciones de Lingüística General[3]; por otro lado, dentro de los proyectos europeos sobre el análisis de la incidencia de las tecnologías informáticas en el aprendizaje o la investigación de las humanidades, la ausencia de universidades españolas y centros de investigación es alarmante.

La única universidad que conozco que ha abierto las puertas a la introducción de la informática humanística en sus programas, con todo derecho, es la Universidad de Deusto (País Vasco), que a partir del curso académico 2003-2004 ofrece el único título oficial en Espa-

[3] En la Universidad Complutense de Madrid (www.ucm.es), llevo tres años impartiendo una asignatura que se acerca a los presupuestos de la *Informática Humanística*: "Informática y textos literarios" (www.ucm.es/info/romanica/informatica.htm); se trata de una asignatura genérica, fuera de los planes de estudio de las catorce licenciaturas que oferta la Faculta de Filología.

ña, al que se le ha denominado "Lenguas Modernas y Nuevas Tecnologías de la Información", que recibirán los alumnos un año antes de terminar su licenciatura, ya sea ésta en Filología Hispánica, Filología Vasca o Filología Inglesa, las tres que oferta la universidad vasca. Los profesores Carmen Isasi y Joseba Abaitua son los responsables de esta iniciativa que ha permitido incluir dentro de las asignaturas de las citadas titulaciones las siguientes materias:

- Nuevas tecnologías en la sociedad de la información[4]
- Aplicación lingüística de las nuevas tecnologías
- Edición digital
- Materiales lingüísticos informatizados
- Gestión de fondos digitales: documentación

El ejemplo de la Universidad de Deusto abre las puertas para ir formando profesionales filólogos que demanda la sociedad; los profesores de lengua y literatura siguen siendo (y seguirán siendo) necesarios, pero, al tiempo, se comienza a solicitar otros perfiles profesionales, ya que se han abierto, gracias a la extensión universal de la "sociedad de la información", otras necesidades; y así, entre las posibles salidas en el mercado laboral de estos nuevos profesionales, se podrían destacar las siguientes, tal y como se indica en el propio portal de la Universidad de Deusto[5]:

1. Documentación en red: empresas de Nuevas Tecnologías, portales de Internet, páginas web.
2. Bibliotecas informatizadas: archivos, bibliotecas públicas y privadas, bibliotecas universitarias, centros de documentación e investigación.

---

[4] El programa de la asignatura, impartido por Joseba Abaitua, así como interesante información bibliográfica, puede consultarse en la actualidad en la siguiente dirección electrónica:
http://www.serv-inf.deusto.es/abaitua/konzeptu/ist0304.htm. Consúltese también el artículo de Carmen Isasi, "Tecnofilología" en la *Revista de la Universidad de Deusto*, enero-marzo, 2003 (http://www.letras.deusto.es/estudios/pn/prensa/03/default.asp?lang=SP).

[5] http://www.letras.deusto.es/estudios/pn/salidas/default.asp?lang=SP.

3. Edición digital: edición digital en medios de comunicación (prensa, radio, televisión), empresa editorial, empresas de Publicidad y Artes Gráficas.
4. Lingüística computacional: fonética acústica, sintaxis y pragmática computacional.
5. Nuevas metodologías en la Enseñanza del Español, y de otras lenguas: enseñanza multimedia, enseñanza a distancia...

3

Pero además de las posibles salidas profesionales –que pueden verse ampliadas si somos capaces de demostrar a la sociedad la necesidad de unos profesionales que dominen a un tiempo la lengua, los textos y las nuevas herramientas y programas informáticos–, la *Informática Humanística* también se presenta pieza fundamental en el diseño y desarrollo de los proyectos científicos que tienen en la acumulación de enormes bancos textuales su razón de ser. Proyectos que se basan en las nuevas tecnologías que permite crear herramientas de recuperación de la información –impensables para las capacidades humanas, como supo ver Busa a finales de los años cuarenta–; pero que sólo podrán ofrecer resultados científicos si tienen en cuenta la naturaleza del texto, las peculiaridades del texto literario, como la filología, en todos sus campos y vertientes, ha mostrado en el último siglo.

En el campo hispánico –como en tantos otros– queda mucho por hacer, aunque también es justo destacar lo que hasta este momento se ha completado con muy buenos resultados, como el proyecto *Philobiblon*, la base bibliográfica de fuentes primarias de textos castellanos, catalanes y gallego-portugueses, coordinado desde la Universidad californiana de Berkeley por Charles Faulhaber (http://sunsite.berkeley.edu/PhiloBiblon/), el portal *Parnaseo* que dirige José Luis Canet desde la Universidad de Valencia (http://parnaseo.uv.es) o los diferentes proyectos que auspicia el Centro Ramón Piñeiro de La Xunta de Galicia (http://airas.cirp.es/); así como merece una especial mención ADMYTE, el *Archivo Digital de Manuscritos y Textos Españoles*, dirigido por Francisco Marcos Marín, cuyo primer CD-Rom vio la luz en 1992 (el segundo en 1993 y el tercero, en 1999). Pero más allá de estas iniciativas, detengámonos con un poco de atención en otro proyecto científico, sin duda uno de los más interesantes y ambiciosos

que tiene como objeto de estudio a los textos hispánicos: el CORDE, una de las bases de datos textuales de la Real Academia Española; proyecto, por otro lado, todavía en pleno proceso de desarrollo y ajuste, que, gracias a la pericia de sus responsables, va mejorando en cada una de sus versiones, aunque, desde el punto de vista de la *Informática Humanística*, se podría incidir en otros aspectos, como intentaré mostrar en las siguientes páginas, que espero que ayuden a concretar algunos problemas a los que hasta ahora no se les ha concedido, desde mi punto de vista, la importancia que merecen, y que matizan y dificultan el uso de los resultados científicos ofrecidos en la actualidad.

El CORDE, acrónimo del *Corpus diacrónico del español*, comenzó su andadura en 1994, y se ha convertido, sin lugar a dudas, en uno de los proyectos pioneros y más ambiciosos del uso de las nuevas tecnologías informáticas en los estudios humanísticos en el ámbito hispánico. En su nueva versión de octubre del 2003, las cifras abruman por su cantidad: 180 millones de registros. ¡Nunca hubo un banco de datos textual similar en cantidad! El CORDE tiene el ambicioso propósito de recoger, de manera sistemática, todos los textos escritos en español, desde sus orígenes hasta 1975, sin discriminar geografías o variantes dialectales o sociolingüísticas.

Los textos escritos alimentan al CORDE (de la misma manera que en el CREA, el *Corpus de Referencia del Español Actual*, que va desde 1975 hasta nuestros días, se utilizan tantos textos escritos como orales, de una gran variedad de procedencias[6]), textos escritos en diferentes épocas y transmitidos en canales tan diversos como el manuscrito o la imprenta moderna, pasando por la imprenta manual y el inédito. Son tres las modalidades de las que proceden los textos, según la propia información del proyecto:

- Libros escaneados a través de un programa de reconocimiento óptico de caracteres (OCR),
- otros conseguidos en formato electrónico,

[6] Tal y como se indica en el portal de Internet de la Real Academia Española: "Los textos escritos, procedentes tanto de libros como de periódicos y revistas, abarcan más de cien materias distintas. La lengua hablada está representada por transcripciones de documentos sonoros, procedentes, en su mayor parte, de la radio y la televisión".

- y algunos tecleados en formato digital, ya que no existía edición moderna de obras que han interesado incluir por la peculiaridad de su lenguaje.

Como ya he tenido ocasión de indicar en otro lugar (Lucía Megías, 2002: 70-75), el CORDE, además de lo abrumador de sus cifras, ofrece una serie de herramientas muy útiles (en principio), para acometer estudios lingüísticos y literarios desde diferentes perspectivas, destacándose la facilidad de uso dado su carácter intuitivo:

a) MOTOR DE BÚSQUEDA. La consulta puede ser libre, o determinada por una serie de factores, que son los que se han tenido en cuenta para diseñar la estructura de la base de datos[7]:

   a.1. Autor
   a.2. Obra
   a.3. Cronológico (desde-hasta)
   a.4. Medio: Libros, periódicos, revistas, miscelánea, orales.
   a.5. Geográfico
   a.6. Tema: lírica, narrativa, teatro, prensa...

b) RECUPERACIÓN DE LA INFORMACIÓN, en varias posibilidades, siguiendo diferentes criterios (casos, autor, año, país, tema y título).

   b.1. Documentos
   b.2. Concordancias
   b.3. Párrafos
   b.4. Agrupaciones, que permiten obtener en un momento un enorme caudal de información para llevar a cabo estudios lingüísticos.

c) ESTADÍSTICAS, que permite conocer cómo un registro se distribuye en el arco diacrónico de su documentación.

d) FILTROS, que permiten afinar mucho más las búsquedas según nuestras necesidades científicas.

---

[7] El tamaño y los criterios de selección, así como el sistema de codificación utilizado puede consultarse en el portal de Internet (www.rae.es), en el *Departamento de Banco de Datos.*

La tecnología informática ha prestado herramientas científicas a la filología hispánica hasta ahora impensables; nunca hasta ahora se habría ni soñado con poder trabajar con un total de millones y millones de registros. Pero, a pesar de los avances tecnológicos, a pesar de las también millonarias inversiones que se han realizado desde 1994, y de las miles de horas que se han utilizado en ponerlo en marcha, ¿por qué el CORDE no se ha convertido en la herramienta científica necesaria, imprescindible para realizar una investigación de lingüística diacrónica o de crítica literaria? ¿Por qué en los congresos científicos de los últimos meses esta herramienta brilla por su ausencia en la mayoría de los casos? Quizás en el diseño y en el desarrollo del proyecto se haya primado el aspecto tecnológico –informático– frente al filológico, cuando serán textos los que albergan y permitan hacer funcionar la tecnología; quizás se haya incidido más en el número de registros que en la naturaleza textual. Quizás este planteamiento sea el adecuado para *corpora* orales, o para los que se nutren de fuentes de información de variadas procedencias, desde el texto literario al periodístico, de las transcripciones de radio y televisión hasta muestras de habla (como el CREA de la propia Real Academia Española); pero en el caso de trabajar con "textos", es necesario –a mi modo de ver– tener en cuenta la problemática específica de su objeto de estudio. La tecnología informática puede compartirse, los métodos y modos de marcación, etiquetado y lematización pueden utilizarse –con algunos ajustes– en ambos proyectos... pero la problemática del objeto de trabajo del CORDE es propia, nueva, particular, característica, por lo que se hace necesario dar respuestas propias, nuevas, particulares y características, como ha venido mostrando la crítica textual en sus (casi) dos siglos de existencia como ciencia. Como también hay que hacerlo teniendo en cuenta el canal de difusión de cada uno ellos: ¿puede tratarse, desde un punto de vista filológico, de la misma manera el *texto* del *Poema de Mio Cid* o *La Colmena* de Camilo José Cela, que las obras completas de Benito Pérez Galdós o los versos de Lope de Vega?

Desde la *Informática Humanística*, la que desde los campos humanísticos –y en este caso, desde el filológico, por lo que deberíamos hablar de *Informática Textual* si queremos ser más precisos– utiliza las nuevas tecnologías para la creación de innovadoras herramientas y de nuevos medios de difusión y de transmisión, son tres los aspectos discu-

tibles que hacen que los resultados del CORDE se tengan que contrastar con otras fuentes, lo que dificulta su uso (en algunos casos) y lo hace poco fiable en otros muchos. Algunos de estos aspectos se explican desde la propia historia de la filología hispánica, que no supo en el siglo XX desarrollar una verdadera ciencia editorial, por lo que hay una carencia de "hipótesis de trabajo científicas" (de ediciones críticas, en otras palabras) que permitan el conocimiento de gran número de textos medievales y renacentistas; pero otros podrían, de manera no muy complicada, corregirse en el actual diseño del proyecto.

3.1. La autoridad

En un momento anterior, se ha hecho alusión a las tres modalidades de las que, según el propio proyecto, proceden los textos: se hablaba de libros escaneados y de otros que se habían conseguido ya en formato electrónico, sin olvidar unos pocos, que por peculiaridades del lenguaje –y también por la mala calidad del papel, como la prensa del siglo XIX– se habían tecleado. Pero, ¿qué ediciones se utilizan? ¿Qué criterio científico se ha utilizado a la hora de escanear, a la hora de aceptar textos digitalizados –como los del proyecto del Hispanic Seminary of Medieval Studies de la Universidad de Madison–, o de teclear algunos de ellos? ¿El criterio es que se trate de una edición accesible, que la haya realizado un académico, que se encuentre en la biblioteca del centro? ¿Qué criterio se ha seguido para descartar una determinada edición, una determinada hipótesis de trabajo de un particular científico? En ningún momento se hacen públicos estos datos, lo que es necesario; pero lo peor es que da la impresión que tampoco en ningún momento se han clarificado en el diseño del proyecto a los criterios que han llevado a los responsables del corpus a elegir una determinada edición por encima de otra. Analicemos un ejemplo concreto para comprender, en todas sus consecuencias, la falta de autoridad de un proyecto cuando se ha primado la cantidad (¡180 millones de registros en la actualidad!) frente a la calidad de los textos –y por *texto* entendemos la última voluntad del autor, a la que sólo se puede acercar uno como hipótesis de trabajo científica, como ha demostrado la crítica textual.

Volvamos a la búsqueda de una determinada forma (*fazienda*), tal y como ya hice en *Filología Románica en Internet. I. Los textos* (Lucía Megías, 2002), cuando se comentaba este mismo problema. Al

buscar *fazienda*, limitando la búsqueda a una época concreta (1200-1500) y al género de la *novela*, se obtienen –¡en sólo unos segundos!– el siguiente resultado: 314 casos en 15 documentos, que son los siguientes, según los datos aportados por el mismo proyecto:

	Casos	Año	Autor	Obra	Tema	Publicación
1	5	1427-1428	Villena, Enrique de	*Traducción y glosas de la Eneida. Libros I-III*	12.Relato extenso novela y otras formas similares	Pedro M. Cátedra, Turner Libros (Madrid), 1994
2	61	1293	Anónimo	*Gran Conquista de Ultramar. Ms. 1187 BNM*	12.Relato extenso novela y otras formas similares	Louis Cooper, Franklin M. Waltman, HSMS (Madison), 1995
3	4	1438	Martínez de Toledo, Alfonso	*Arcipreste de Talavera (Corbacho)*	12.Otras formas	Marcella Ciceri, Espasa-Calpe (Madrid), 1990
4	4	1300-1325	Anónimo	*Cuento muy fermoso de Otas de Roma*	12.Relato extenso novela y otras formas similares	Herbert L. Baird, Jr., RAE (Madrid), 1976
5	5	1499	Anónimo	*La historia de los nobles caballeros Oliveros de Castilla y Artús d'Algarbe*	12.Relato extenso novela y otras formas similares	Nieves Baranda, Turner Libros (Madrid), 1995
6	47	1313-1498	Anónimo	*El baladro del sabio Merlín con sus profecías*	12.Relato extenso novela y otras formas similares	Isabel Hernández González, CILUS (Salamanca), 1999
7	48	1325-1335	Manuel, Juan	*El Conde Lucanor*	12.Relato breve culto	Guillermo Serés, Crítica (Barcelona), 1994

11	5	1300-1305	Anónimo	*Libro del cavallero Cifar*	12.Relato extenso novela y otras formas similares	Juan Manuel Cacho Blecua, Universidad de Zaragoza (Zaragoza), 2003
12	86	1251	Anónimo	*Calila e Dimna*	12.Breve	Juan Manuel Cacho Blecua; María Jesús Lacarra, Castalia (Madrid), 1993
13	2	1425-1450	Rodríguez del Padrón, Juan	*Bursario*	12.Relato extenso novela y otras formas similares	Pilar Saquero Suárez-Somonte; Tomás González Rolán, Universidad Complutense (Madrid), 1984
14	15	1482-1492	Rodríguez de Montalvo, Garci	*Amadís de Gaula II*	12.Relato extenso novela y otras formas similares	Juan Manuel Cacho Blecua, Cátedra (Madrid), 1991
15	20	1482-1492	Rodríguez de Montalvo, Garci	*Amadís de Gaula, libros I y II*	12.Relato extenso novela y otras formas similares	Juan Manuel Cacho Blecua, Cátedra (Madrid), 1991

En este listado de quince obras, se documenta una alarmante heterogeneidad de ediciones, tanto por sus características como por sus finalidades, que hace imposible –o al menos así nos lo enseña la filología– la comparación de sus resultados, lo que sí se espera al incluirlas todas en un mismo banco de datos, sin ningún criterio diferenciador entre ellos –lo que sí que permitiría la tecnología informática, por otro lado. De este modo, encontramos:

a) Las transcripciones de un determinado testimonio, como así sucede con los procedentes del proyecto del Hispanic Seminary of Medieval Studies de Madison, en donde se seguía un particular sistema de transcripción, pensado para su posterior uso informático con la finalidad de realizar un diccionario del español medieval, y nunca para su lectura como libro, a pesar de que hayan sido

muchos los textos que sólo están disponibles en un acercamiento editorial moderno gracias a esta colección. Dentro de este grupo se engloba el nº 2: la transcripción del ms. 1187 de la Biblioteca Nacional de Madrid, que ha conservado sólo los últimos ciento treinta y cinco capítulos del tercer libro y el cuarto completo de la *Gran Conquista de Ultramar* (DFLME, 2002: 603-608).

b) Ediciones modernas, que se han publicado en colecciones donde se prima la transcripción crítica de un testimonio manuscrito, como es la Biblioteca Castro: nº 1, 5 y 9; o en nº 6, que es también una transcripción del incunable de uno de los primeros textos caballerescos.
c) Ediciones modernas, publicadas en editoriales y colecciones de prestigio (como Castalia, Cátedra, Crítica o Espasa-Calpe), que ofrecen una gran variedad de posibilidades de acercamiento textual, desde los más apegados a un determinado testimonio, a los que se acercan al estudio completo de su transmisión: nº 3, 7, 14 y 15 (en donde no entiendo muy bien la diferencia de estos dos últimos, ya que se trata de la misma edición de la obra).
d) Ediciones críticas, en donde los investigadores ofrecen una determinada hipótesis de trabajo, después de analizar la transmisión completa de una determinada obra: nº 13.
e) Ediciones realizadas a finales del siglo XIX o principios del siglo XX (nº 8 y 10), con unos criterios científicos muy alejados –en ocasiones– de los mínimos presupuestos ecdóticos actuales.

Lo cierto es que la ausencia de una tradición ecdótica, de la escuela que sea, en España, dificulta, en gran medida, el diseño y el desarrollo de proyectos textuales como el presente, de una ambición desbordante, lo que no sucede en otros países como Italia o Francia, al margen de que en un caso nos movamos en la línea certera del lachmannismo y del neo-lachmannismo, y en el otro en el arte propugnado por Bédier, que ha terminado por primar la transcripción crítica de testimonios. Pero la dificultad de partida –enorme, repito– no justifica la mezcla de "autoridad" a la hora de trabajar con ediciones: no es suficiente con haber escaneado bien un libro, con haberlo tecleado adecuadamente, con haberlo marcado y etiquetado de manera correcta, con haber conseguido que los programas informáticos lo reconozcan –labor que compete, sobre todo, a los técnicos informáticos–; es necesario,

además, que estos libros hayan sido seleccionados, o hayan sido transcritos o editados, de acuerdo a unos criterios filológicos, que son los que marcarán su "autoridad", como la *Informática Textual* ha venido enseñando –y lo seguirá haciendo– en los últimos años.

Del listado anterior hemos dejado fuera una de las entradas, que merece un comentario final, por ser uno de los aspectos novedosos que se han introducido en la última versión del CORDE, de octubre del 2003: el nº 11, que se corresponde con el texto del *Libro del cavallero Zifar*, con la siguiente indicación: "Juan Manuel Cacho Blecua, Universidad de Zaragoza (Zaragoza), 2003". En la presentación que en Internet se hace de esta nueva versión, se destaca una de las mejoras que se han introducido, que se relaciona directamente con lo que vamos analizando hasta aquí: "Además del considerable incremento del volumen de textos, la nueva versión ha sustituido algunas ediciones por otras más adaptadas a los niveles de calidad habituales en las ediciones actuales". En la anterior versión, como ya se indicó en otra publicación (Lucía Megías, 2002: 75), se utilizó para introducir en el CORDE el texto del *Zifar* (principios del siglo XIV), la transcripción que Francisco Gago Jover había realizado de un testimonio manuscrito de finales del siglo XV (ms. P de la Bibliothèque Nationale de France), para el HSMS de Madison. La sustitución por una edición de Juan Manuel Cacho Blecua no puede ser más acertada: pocos investigadores pueden encontrarse hoy que conozcan mejor el texto del *Zifar*. Pero, ¿qué criterios ha utilizado Juan Manuel Cacho Blecua para su edición? ¿La transcripción de uno de los tres testimonios conservados? ¿La edición crítica del texto? En el caso de las ediciones publicadas, siempre existe la posibilidad de consultarlas en la biblioteca, pero ¿qué sucede con estos "textos", que nacen del acuerdo de colaboración entre la Real Academia Española y diferentes universidades y centros científicos de España? No me cabe duda de que se ha ganado en calidad textual, pero no así en el aspecto de "autoridad" que aquí se viene comentando.

Una posible solución a este problema pasa por hacer público, al tiempo que la ficha mínima del "documento", como se denomina en el proyecto, dos tipos de informaciones:

a) El listado completo de los textos que, en cada momento, están incorporados al CORDE, con la referencia bibliográfica pertinente, ya sea la edición de la que se ha escaneado, ya sea el proyecto

científico que ha presentado el texto ya digitalizado. Si un usuario desea buscar una forma en un autor concreto, en un determinado texto, por ejemplo el *Cantar de Mío Cid* o en una obra particular de Benito Pérez Galdós, podría conocer, de antemano, si ésta ha sido incluida.

b) Por otro lado, y dada esa carencia filológica de la ciencia hispánica durante el siglo XX, de la que ya nos hemos lamentado en estas páginas, se haría necesario acompañar este mínimo listado –que ya existe, como se aprecia en los datos antes expuestos de los resultados de la búsqueda de *fazienda*– de unas notas filológicas, en donde se expusieran, de una manera clara y concisa, [1] las características de la edición utilizada como base, [2] la problemática textual que rodea al texto –siempre que se considere oportuno–, y [3] la clasificación de la edición utilizada, teniendo en cuenta una tipología previamente establecida, que va desde la edición crítica a la mera transcripción paleográfica de un determinado testimonio.

De esta manera, podríamos –desde el propio proyecto– ofrecer al usuario los datos necesarios para que pueda contrastar los resultados científicos obtenidos. No se trata –¡por supuesto!– de realizar ediciones nuevas de todos los textos sino de todo lo contrario: de utilizar lo que se ha hecho, de mejorarlo y de situarlo en su verdadero contexto científico y filológico. Una vez más, se muestra la necesidad de tener en cuenta la *Informática Textual;* una vez más, se muestra la necesidad de tener en cuenta las peculiares características de los *textos* a la hora de dar forma a las nuevas herramientas informáticas.

3.2. La presentación gráfica

El diverso origen de las diferentes ediciones y transcripciones de las que se vale el proyecto para llegar a esas escalofriantes cifras de millones de registros, tiene también una consecuencia que limita la capacidad de búsqueda: la heterogeneidad de las presentaciones gráficas. La búsqueda de *fazienda* (entre 1200-1500 en obras narrativas) nos daba el resultado de 314 casos en 15 documentos... pero no es la única forma de esta palabra en el corpus: *façienda* aparece documentada una vez en la misma transcripción que Louis Cooper y Franklin M. Waltman realizaron para el Hispanic Seminary of Medieval Studies

(Madison) en 1995, del ms. 1187 de la Biblioteca Nacional de Madrid, que ha transmitido *La Gran Conquista de Ultramar*, que documenta 61 casos de *fazienda*.

Y así lo podríamos multiplicar con otros ejemplos, como sucede con el infinitivo *fazer*, que da resultados diferentes según la presentación gráfica, entre 1200-1500, en textos narrativos:

a) fazer: 2978 casos en 24 documentos.
b) facer: 7 casos en 2 documentos.
c) ffazer: 1 caso en 1 documento.

Este aspecto, que limita las enormes posibilidades que ofrece la tecnología informática para poder trabajar con cantidades tan abundantes de información, podría corregirse mediante dos procedimientos: adaptar a unas mismas normas de presentación gráfica todo el corpus (organizado, quizás, por épocas), como así lo realizó el proyecto del HSMS, que sí que ha sido volcado con sus modelos en el CORDE; o, lo que es preferible, en vez de trabajar con formas hacerlo con lemas, en donde se recogen todas las formas que lo han documentado, en sus particulares modos de presentación, lo que permitiría estudios grafemáticos y lingüísticos, que ahora no pueden realizarse con el proyecto en su formulación actual. La lematización y la etiquetación de los *corpora textuales* se ha convertido en uno de los requisitos imprescindibles de este tipo de proyectos. Al mismo tiempo, se podría incluir, además de las ediciones, transcripciones paleográficas de los testimonios, lo que permitiría avanzar en el estudio de la lingüística diacrónica.

3.3. La crítica textual

Por último, hay un error de base –lección de la crítica textual– que limita, en gran medida, el uso que podemos hacer de algunas de las herramientas informáticas que ofrece el CORDE: la diferencia entre *texto* y *testimonio*, entre *génesis* y *transmisión*. Este error hace inútiles, en gran medida, los datos que aparecen en herramientas como las estadísticas, que se basan en el marcado cronológico que se ha hecho de cada texto... y este marcado cronológico (que aparece en una de las primeras columnas del listado anterior) se ha realizado sobre la génesis del texto, y no sobre la datación de los testimonios que se toman como texto base

de las diferentes transcripciones o ediciones que se han utilizado en el proyecto. Veamos unos pocos ejemplos, siguiente el listado anterior:

> a) nº 6: *El Baladro del sabio Merlín*: que se data con la fecha doble: 1313-1498: la primera corresponde –de manera hipotética– a la datación de su traducción al castellano (más bien, la de un testimonio manuscrito conservado en la Biblioteca Universitaria de Salamanca, el ms. 1877); mientras que la segunda es la fecha del incunable, que sirve de base a la transcripción de Isabel Hernández. ¿Por qué no marcar de alguna manera esta información? ¿Por que no diferenciar claramente lo que es la documentación del testimonio (1498) de lo que es el texto original, que se ha perdido, y cuya forma lingüística nada tendrá que ver con el incunable?
>
> b) nº 8: *La estoria de Merlín*: que sucede otro tanto: la fecha de 1313 vuelve a hacer alusión a la génesis, pero el testimonio manuscrito –una verdadera compilación sapiencial, por otro lado– está fechado en 1469, que es a la fecha que deberíamos adscribir, en última instancia, la forma lingüística del texto tal y como aparece en el testimonio conservado.
>
> c) nº 12: *Calila e Dimna*, que se data en 1251; la edición de Juan Manuel Cacho Blecua y de Mª Jesús Lacarra utiliza como texto base el ms. A, el que se conserva en la Biblioteca del Monasterio de El Escorial: ms. h-III-9, datado a principios del siglo XV. Estas son las palabras de los editores a la hora de justificar la elección de A como texto base: "De los medios externos se desprende que el manuscrito A es el más antiguo, y de un cotejo interno se puede deducir que el manuscrito A cuantitativamente está menos modernizado. Los resultados no son concluyentes por cuanto hay ciertas formas que en el manuscrito A, ocasionalmente, resultan más modernas. Por poner solo un caso, hay formas en –ié del imperfecto que se conservan en B y han desaparecido en A. Resulta sorprendente que un copista medieval hubiera mantenido del texto sin modernizarlo, cuando para él suele ser algo vivo y susceptible de adaptarse a los nuevos tiempos" (1993: 65). ¿Podemos seguir atribuyendo determinado léxico, determinadas formas del *Calila e Dimna* del manuscrito del siglo XV al 1251, fecha de su posible génesis?

d) nºs 14 y 15: *Amadís de Gaula*: que se data según el texto refundido de Garci Rodríguez de Montalvo, o de su perdida primera edición de 1482-1492, cuando lo que hemos conservado –y lo que se edita– es la edición de 1508 que viera a la luz en los talleres zaragozanos de Jorge Coci. ¿Debemos incluir *Amadís de Gaula* dentro de nuestra pesquisa entre 1200-1500, cuando las primeras documentaciones impresas se datan en 1508?

Veamos ahora los resultados que aparecen en la herramienta de "estadísticas" según la búsqueda de *fazienda* con los criterios anteriormente indicados:

Estadísticas

Año	**País**	**Tema**
%	**%**	**%**
Casos	**Casos**	**Casos**
1251	*ESPAÑA*	*11.- Prosa lírica*
27.38	100.00	100.00
86	314	314
1293		
19.42		
61		
1313		
16.24		
51		
1325		
15.28		
48		
1482		
11.14		
35		
1300		
2.86		
9		

1498
2.22
7

1427
1.59
5

1499
1.59
5

Otros
2.22
7

¿Debemos seguir adscribiendo, como ya se ha indicado, a 1251 los resultados del ms. del siglo XV del *Calila e Dimna*? ¿Por qué se toma la fecha de 1313 para los textos sobre la *historia de Merlín*, cuando pertenecen a un incunable de 1498 y a una copia de finales de la misma centuria? ¿Acaso las documentaciones (35) de 1482, pertenecientes al *Amadís de Gaula*, no deberíamos situarlas con todo rigor en 1508? Y las 9 documentaciones del *Libro del cavallero Zifar*, que se datan en 1300, ¿pertenecen a un de los dos testimonios manuscritos del siglo XV o al impreso de 1512?

La distinción entre *texto* y *testimonio*, entre génesis y transmisión, la marcación adecuada de cada una de las fuentes de información del proyecto desde la perspectiva de la *Informática Humanística*, podría solucionar este problema que limita, en gran medida, la utilización científica de las herramientas informáticas que pone a disposición de la comunidad científica el CORDE. Las etiquetas de los diferentes "textos" incluidos estaría estrechamente relacionada con las notas filológicas que deberían acompañar a los "documentos", tal y como se ha indicado con anterioridad. De este modo, se podría –de acuerdo con el análisis de las ediciones utilizadas– realizar un doble etiquetado: el del texto (1251 en el caso del *Calila e Dimna*) y el del testimonio que se toma como base (siglo XV) o el que se transcribe exclusivamente. La actualización lingüística de las copias a lo largo de la Edad Media y de la difusión por

medio de la imprenta manual (hasta el siglo XIX) es una realidad evidente para todos los que hemos tenido que enfrentarnos a una edición científica. Los sueños positivistas del siglo XIX de "reconstrucción lingüística" del original han quedado en eso: en pesadillas editoriales. Por otro lado, gracias a la distinción clara entre *texto* y *testimonio*, se podrían incluir dentro del proyecto las transcripciones de varios testimonios de un mismo texto, con nuevas posibilidades de herramientas informáticas (y de modificación de algunas existentes) para su uso científico. Y este punto puede relacionarse con el anterior, para permitir también la introducción de transcripciones paleográficas (en especial de documentos) que permitieran estudios grafemáticos y de fonético y fonología histórica, por sólo poner unos ejemplos.

El CORDE, como ya se ha indicado, es un proyecto científico en marcha, en el que se lleva trabajando desde 1994, pero que todavía le queda mucho esfuerzo por delante. ¿Qué es lo que se ha conseguido hasta ahora? Sus 180 millones de registros constituyen el mayor almacén de ediciones y de transcripciones que hasta ahora se había podido soñar; el mayor banco de datos textuales de la historia de la filología hispánica. Pero al primarse –en gran medida– la cantidad, el número de este almacén, se ha desatendido –o al menos, así se percibe como usuario– el aspecto filológico, lo que podrá corregirse –si así se considera oportuno– en los próximos años. ¿Acaso la tecnología informática no ofrece más posibilidades si sabemos aunarla con los conocimientos humanísticos, como nos enseña la *Informática Textual*, y no quedarnos simplemente en la acumulación de la *información*, que nunca se ha de confundir con *conocimiento*?

4

Desde el ámbito hispánico, como hemos querido mostrar en sólo estos dos ejemplos, uno académico y otro de investigación, la *Informática Humanística* está todavía en mantillas, a demasiada distancia de lo que se está investigando y enseñando en otros países. Los proyectos, tanto privados como públicos, que tienen a los textos y a la informática como componentes esenciales (bibliotecas virtuales, banco de datos, etc.) cada vez son más numerosos; pero en todos ellos, existe una carencia de filología que termina por hacer poco productivos –nulos desde el punto de vista científico– sus resultados. La *Informática Humanística*, desde una

de sus vertientes, la *Informática Textual*, vendría a dar una respuesta a los nuevos problemas científicos y académicos que están surgiendo a principios del siglo XXI, que están demandando una respuesta inmediata.

El 16 de junio del 2003, unos ciento treinta profesores italianos y extranjeros escribieron una carta al Ministro de Educación italiano, solicitándole la creación de un área de conocimiento con el título de "Informatica Umanistica", en contestación a diferentes informaciones aparecidas en la prensa por aquellos días[8]. Esta carta es el resultado de varios años de trabajo, que ha dado ya lugar a una amplia polémica y bibliografía, así como a las primeras licenciaturas y cursos específicos en diversas universidades[9].

En enero del 2003, un grupo de 18 profesores de diferentes universidades italianas, había realizado una propuesta concreta, que, por su importancia y para aclarar los límites y las competencias de esta nueva área de conocimiento, traduzco a continuación[10]:

> Formulamos la propuesta de constitución de un Sector científico-disciplinar, denominado INFORMATICA APLICADA A LAS DISCIPLINAS

---

[8] La carta completa, así como el listado de adhesiones –que en octubre del 2003 supera los 170-, puede consultarse en la siguiente dirección en Internet: http://193.205.145.117/docenti/informatica/appello/index.htm.

[9] Un panorama de las críticas y de las posibilidades de la *Informática Humanística* en Italia puede consultarse en los diferentes artículos recogidos en la revista electrónica *Griseldaonline*: http://www.griseldaonline.it/informatica/index.htm, en donde el lector interesado encontrará diversas referencias bibliográficas y temáticas.

[10] El original italiano puede consultarse en http://193.205.145.117/docenti/informatica/appello/gruppocun_2.pdf; así como en formato web en la revista Griseldaonline, firmado por Tito Orlandi, el promotor de la misma: http://www.griseldaonline.it/informatica/orlandi.htm. Los trabajos y publicaciones del profesor Orlandi, director del Centro Interdipartimentale per l'Automazione delle Discipline Umanistiche (CISADU), sobre este tema son numerosos, y muchos de ellos pueden consultarse en su portal personal: http://rmcisadu.let.uniroma1.it/~orlandi/. Consúltese, también, Computing In Humanities Education A European Perspective, Edited by Koenraad de Smedt Hazel Gardiner, Espen Ore, Tito Orlandi, Harold Short, Jacques Souillot, William Vaughan (1999: http://helmer.hit.uib.no/AcoHum/book/).

HUMANÍSTICAS (INFORMÁTICA HUMANÍSTICA), que debe incluirse en el Área 10: Ciencias de la antigüedad, filológica-literarias e histórica-artísticas, y en el Área 11: Ciencias históricas, filosóficas, pedagógicas y sicológicas, con la siguiente declaración:

El sector comprende las competencias relativas a las innovaciones metodológicas producidas por la utilización de los medios informáticos y por los sistemas multimedia en el ámbito de las disciplinas humanísticas, en especial en cuanto atañe a la representación de datos, la formulación de diversas fases de trabajos en la investigación, y la técnica de difusión de sus resultados.

En particular, en el ámbito de los *Estudios Lingüísticos* toma en consideración los problemas relacionados con el estudio de los formalismos del lenguaje y las técnicas de análisis de las reglas y de las estructuras lingüísticas y léxicas, de generación automática de enunciados lingüísticos, del diseño, gestión y difusión en línea de repertorios y bases de datos lingüísticos, léxicos y terminológicos ("industria de la lengua").

En el ámbito de los *Estudios Filológicos*, toma en consideración el trabajo de crítica y ecdótica, basado en la gestión automática de manuscritos codificados y registrados en soporte magnético.

En el ámbito de los *Estudios Históricos* se ocupa del análisis y de la síntesis de la documentación reunida en los bancos de datos históricos, y su difusión multimedia.

En el ámbito de los *Estudios Filosóficos e Histórico-Filosóficos* se ocupa del tratamiento automático de textos, del análisis de los lenguajes filosóficos, del alcance teórico de los diferentes modos de representación de la conciencia, y de la posibilidad de traducir problemáticas filosóficas en las nuevas metodologías, formas de expresión, y modalidad de comunicación.

En el ámbito de los *Estudios Arqueológicos* se ocupa de las nuevas metodologías para el análisis y la importancia del territorio, y para la gestión y la representación de los datos de las excavaciones. Se ocupa también de los instrumentos para los inventarios, la catalogación y la clasificación de los bienes arqueológicos, mediante la creación de bases de datos y el uso de métodos estadísticos.

En el ámbito de los *Estudios Literarios* se ocupa de los nuevos métodos de investigación y de análisis de los textos con herramientas multimedia.

> En el ámbito de los *Estudios Histórico-Artísticos* se ocupa de los nuevos criterios de estimación de las obras de arte, mediante el análisis numérico de los componentes gráficos y pictóricos, y de los métodos multimedia para la catalogación, la conservación y el disfrute del patrimonio histórico-artístico.
>
> En el ámbito de los *Estudios Históricos-Musicales* se ocupa del examen de las partituras con la finalidad de la reconstrucción histórica y del análisis musical.
>
> En el ámbito de las *Metodologías Didácticas* se ocupa del uso correcto de los instrumentos computacionales en el diseño y en el desarrollo de la actividad didáctica.

La cita ha sido extensa y marca la ambición del proyecto universitario y académico que en estos momentos se está defendiendo desde Italia, desde una ciencia y una filología muy cercana a la que se ha extendido en el mundo hispánico. Un proyecto que entiende los estudios humanísticos desde una visión muy amplia, ya que, con diferentes posibilidades y varias realidades, en todos los ámbitos indicados las nuevas tecnologías informáticas han venido –lo queramos o no– no sólo a modificar en gran medida la comunicación y la difusión de los resultados científicos, sino también las propias bases de su metodología.

Estas páginas volanderas escritas desde el asombro al ver cómo también en este campo el ámbito hispánico parece estar condenado a un retraso frente a otros países, no quieren agotar el tema, sino todo lo contrario. Nuestro planteamiento no se relaciona tanto con el académico y universitario que viene a documentar el caso italiano, sino que mira más a la necesidad de ir creando estructuras de trabajo, interdisciplinares, en donde se den respuestas a nuevos canales de difusión, de enseñanza, pero también de creación textual, que permiten las nuevas tecnologías, y que cada día van a ser más numerosas y más complejas. En este sentido, dentro de la *Informática Humanística*, que ha de ser amplia en sus límites, interesa, por último, destacar la necesidad de crear un ámbito específico relacionado con el texto, que se podría denominar *Informática Textual*, que debe prestar especial atención a todas las cuestiones que rodean al texto, desde su génesis, difusión, análisis y didáctica; por lo que dentro de esta rama, podríamos hablar de cuatro ámbitos de trabajo:

1. DOCUMENTAL: diseño y gestión de bases de datos y de archivos hipertextuales.
2. EDITORIAL: diseño del hipertexto, de las ediciones hipertextuales, de la estructura de las bibliotecas virtuales, difusión de hipertextos informativos e, incluso, de hipertextos creativos.
3. INSTRUMENTAL: uso y perfeccionamiento de programas específicos para el análisis de los textos.
4. DIDÁCTICO: diseño de nuevas modalidades de enseñanza gracias al hipertexto y a las estructuras hipertextuales.

Las nuevas tecnologías no sólo han venido a facilitarnos una comunicación más cómoda y rápida (correo electrónico), unos medios nuevos de edición y de escritura (imprenta digital, programas de tratamiento de texto), sino que pone a disposición nuevas herramientas de trabajo científico y nuevas modalidades de difusión (hipertexto), que plantean diversos retos y preguntas, a las que deberemos dar respuesta en los próximos años; respuestas que pasan por aunar esfuerzos en una misma disciplina, que no puede ser otra que la *Informática Humanística* (y la *Informática Textual* para nuestro ámbito más específico). No se trata de contar con más información, con más datos –aunque estos sean millonarios–, de morir ahogados –e incomunicados– entre tanta información, sino de disponer de la metodología, las herramientas, la estructura, la metodología adecuada para convertir esta *información* en *conocimiento*, el verdadero reto de nuestras disciplinas humanísticas.

Referencias bibliográficas y sitográficas

DFLME (Diccionario Filológico de la Literatura Medieval Española), dirigido por Carlos Alvar y José Manuel Lucía Megías, Madrid, Castalia, 2002.

GIGLIOZZI, G., 1997. *Il testo e il computer*, Mondadori, Milano.

——, 2002. *La fondazione dell'informatica applicata al testo letterario*, a cura di Raul Mordenti, número especial de la serie "Testo e Senso", Editrice Universitaria di Roma.

GRUBER, Daniela y Patrick PAULETTO, *Umanesimo & Informatica, Le nuove frontiere della ricerca e della didattica nel campo degli studi letterari,*

Atti del Convegno, Trento 24-25 maggio 1996, a cura di Daniela e, puede consultarse en PDF en Internet: http://circe.lett.unitn.it/circe/html/attivita/uman.asp

LEONARDI, Claudio - MORELLI, Marcello - SANTI, Francesco (cur.), 1994. *Macchine per leggere. Tradizioni e nuove tecnologie per comprendere i testi* (Collana della fondazione Ezio Franceschini), Spoleto, Centro italiano studi sull'alto medioevo.

LUCÍA MEGÍAS, José Manuel, 2002. *Literatura románica en Internet. I. Los textos,* Madrid, Castalia.

——, 2003. "La crítica textual ante el siglo XXI: la primacía del texto", en Lillian von der Walde Moheno (ed.), *Propuestas teórico metodológicas para el estudio de la literatura hispánica medieval,* México, Universidad Autónoma Metropolitana Iztapalapa-editorial Plaza y Valdes, pp. 417-90.

NEROZZI BELLMAN, P. (ed.), 1997. *Internet e le Muse,* Mimesis, Milano.

ORLANDI, Tito, 1990. *Informatica umanistica,* Nuova Italia Scientifica, Roma.

——, 1990. *Per l'Informatica nella Facoltà di Lettere,* Roma, Bulzoni.

—— "Proposta: Informatica applicata alle discipline umanistiche (ovvero: Informatica umanistica", *Griseldaonline* (http://www.griseldaonline.it/informatica/index.htm): octubre del 2003.

RICCIARDI, Mario (cur.), 1995. *Scrivere comunicare apprendere con le nuove tecnologie* (I nuovi strumenti del sapere umanistico), Torino, Bollati Boringhieri.