Revista de Humanidades Digitales

Para unas lecturas remediadas: análisis cuantitativo y cualitativo de textos*

For Some Remediated Readings: Quantitative and Qualitative Texts Analysis

Amelia SANZ CABRERIZO
Universidad Complutense de Madrid
amsanz@ucm.es
https://orcid.org/0000-0002-1654-5000

Todo esto no es nuevo, esto ya no es innovación: vivimos en un mundo, en este Occidente, cuando las pantallas grandes y pequeñas se han hecho omnipresentes, después de la revolución digital y con ella. Todos somos, en una palabra, postdigitales.

Y es una época y un mundo que parece tener más en cuenta los números que las letras. Será porque la modernidad siempre concedió una importancia fundamental a las matemáticas: desde Galileo, cuando afirmó que el universo está escrito en caracteres matemáticos, y desde Descartes, con su *mathesis universalis*. Frente a sofistas, nominalistas, postmodernos y todos los que han señalado la fluctuación de las palabras, las cifras parece que se han convertido en garantía de uso. Hoy los ordenadores no saben de la naturaleza de lo que cuentan, pero cuentan: no saben qué, pero sí cuántos.

Así nos vemos sumidos en un proceso de aculturación (hacia la cultura digital) que afecta de lleno a nuestros estudiantes, a sus modos y recorridos de aprendizaje, a sus desarrollos profesionales, personales. Más aún, convergen diferentes tecnologías que estaban separadas: escribir, almacenar y preservar, transferir de un medio a otro, publicar y comunicar, todo se reúne en una máquina y en una máquina de máquinas y en otras aún mayores y distribuidas. De esta forma, las tecnologías de la escritura y de la lectura se han convertido en una poderosísima maquinaria de exclusión social: poder ser con o sin pasaporte digital.

Y es que nunca hemos leído y escrito tanto como ahora. Así nos encontramos con otras formas de escritura (como la tuiteratura) y de lectura (como las redes Goodreads y Babelio[1]) que exigen dispositivos electrónicos para acceder a ellos, también para recuperarlas y analizarlas. No son textos, son textualidades y son electrónicas: entramados multimediales de comunicación que pueden ser leídos, oídos, vistos y permiten diferentes modos de interacción con en-

RHD 6 (2021)
ISSN
2531-1786

Recibido 20/05/2021 – Aceptado 13/07/2021

laces, pasos de página, descargas, comentarios. Todo ello genera una cantidad sin precedentes de materiales escritos que podemos conservar o perder, pero que, en cualquier caso, nosotros (los ciudadanos) o ellos (GAFAM[2]) habremos de decidir qué va a quedar como memoria de esta época, en qué espacios de almacenamiento, qué queremos hacer con ellos, a qué precio.

Y es un hecho que los dispositivos electrónicos han transformado nuestra forma de escribir y de editar un texto, mientras que cierta lectura parece resistirse: para muchos la lectura sigue siendo más cómoda en el viejo medio impreso y las grandes bibliotecas digitales son más bien escaparates a los que pegamos manos, nariz y ojos sin que podamos alcanzar y apropiarnos de los objetos.

Pero es que se trata de otro sistema de representación: si el emisor y el receptor de una carta compartían y conocían la tecnología de su interfaz (la hoja de papel, la tinta y sus signos escritos), ahora entre autor y lector hay unos códigos (la lógica artificial de los lenguajes de programación) que ambos desconocen, pero cuya mediación aceptan. Las grandes compañías ofrecen sus productos cerrados, cual cajas negras inexpugnables, que los ciudadanos tomamos o quedamos excluidos.

Por todo eso necesitamos una alfabetización digital masiva, transversal, continua, reglada. Este monográfico dedicado al análisis cuantitativo y cualitativo de textos prueba que es posible pasar de la información a la formación en la universidad cuando de alfabetización digital se trata. Es fruto de los talleres que, cada primavera, la Facultad de Filología de la Universidad Complutense propone a sus doctorandos[3]. Los autores de estos artículos son doctorandos o ya doctorados que han sabido apropiarse de las herramientas propuestas para contestar a sus propias preguntas de investigación y medir así el alcance de la metodología digital.

Y es que se trata de un deber de la enseñanza superior: no porque el manejo de un paquete determinado de herramientas sea la llave de un empleo fácil, sino porque estamos obligados a trabajar competencias y habilidades y, con ellas, identidades y valores, esto es, esa metaconciencia crítica de cambio cultural, esencia misma de la universidad.

No es tan difícil: en las facultades de Letras, y de Humanidades en general, somos expertos en etiquetado, en categorizaciones y catalogaciones con los lenguajes naturales. Tenemos que entender y practicar cómo se organiza la información en los lenguajes de programación, tan empeñados en eliminar ambigüedades y pluralidades. No es necesario un genio particular, solo curiosidad, un poco de dedicación y ganas de resolver problemas: podemos humanizar la informática, podemos feminizar la informática, podemos robar el fuego porque también es nuestro.

* Este monográfico se gestó dentro del del Proyecto de Investigación REC-LIT. *Reciclajes culturales: transliteraturas en la era postdigital* (Referencia RTI2018-094607-B-I00), financiado por FEDER/Ministerio de Ciencia e Innovación – Agencia Estatal de Investigación .

[1] Accesibles desde https://www.goodreads.com/ y https://www.babelio.com/.

[2] GAFAM es el acrónimo de Google, Amazon, Facebook, Apple y Microsoft.

[3] Por ejemplo, Herramientas digitales para el análisis cuantitativo y cualitativo de textos (2020). Más detalles en: https://eventos.ucm.es/46888/detail/herramientas-digitales-para-el-analisis-cualitativo-y-cuantitativo-de-textos-digital-tools-for-qua.html.

Es fundamental que las competencias y herramientas digitales entren de manera transversal en todas las asignaturas y seminarios sobre lenguas, sus culturas y sus literaturas. Para ello, damos pasos como el que mostramos en esta publicación en favor de una propuesta curricular propia en la que los estudiantes puedan aprender métodos y funcionalidades para una lectura atenta y reflexiva, tanto a gran escala como a micro-escala, con lentes digitales. Porque ya no es el momento de aislar en una sola asignatura todos los principios y funcionalidades propios de las llamadas Digital Humanities: ya están en todas partes, ya no son una especialidad, sino una necesidad general, incluso en España donde la financiación para estas actividades es mínima.

Es tiempo de elegir herramientas y métodos para responder a necesidades y preguntas precisas en un grado o en un máster, en cualquier asignatura o seminario: accesibles a todos, sin necesidad de adaptación especial, flexibles según las elecciones teóricas y metodológicas de cualquiera, simples e intuitivas. Serían unas humanidades digitales *blandas* que abrirían la puerta fácilmente a los procelosos senderos del machine learning… Y podemos demostrar que no es necesario un gran equipamiento ni siquiera un aula especial, sino que basta con los simples portátiles y una conexión en casa; tampoco necesita el profesor ser un experto de la computación, sino elegir unos conceptos y una pregunta, un corpus, unas herramientas y las habilidades que todo ello conlleva y se van a desarrollar; ni siquiera los estudiantes, que no se atreven a confesar el miedo a la máquina, se van a sorprender si la metodología y las herramientas están perfectamente engastadas en el programa de una asignatura: hay instrumentos que nos permiten desarrollar un ejercicio en un cuarto de hora (como las nubes de palabras) o en media hora (la exploración de una biblioteca digital); otros pueden llevarnos una clase entera y convertirse en un auténtico acontecimiento (crear una entrada en Wikipedia) o toda una semana (la anotación colaborativa de un texto).

Pero tenemos varios obstáculos: la dificultad para constituir los corpus, la dificultad de conocer al día las herramientas disponibles.

Efectivamente, la promesa del todo digitalizado no se está cumpliendo: Europeana confiesa que solo el 10% del patrimonio europeo ha sido digitalizado, Enumerate anuncia que el 30% de ese patrimonio no será digitalizado[4]. Podemos admirar los recursos y corpus a disposición del lingüista[5], pero lo cierto es que, en los estudios literarios e históricos en general, tenemos grandes problemas: prueben constituir un corpus de millones de palabras con textos clásicos a partir de Perseus, Scaife Viewer o Lasla[6], intenten descargar en modo texto toda Galiciana[7], atrévanse a lanzar cualquier aseveración sobre la literaturas anglófonas a partir del corpus de Chadwick[8] o sobre las literaturas hispanas gracias a la Biblioteca Virtual Cervantes[9]. La disponibilidad de los

[4] ENUMERATE. Survey Report on Digitisation in European Cultural Heritage Institution (accesible desde: http://enumeratedataplatform.digibis.com/reports/survey-report-on-digitisation-in-european-cultural-heritage-institutions-2015/detail).
[5] CLARIN, Ressources families (accesible desde : https://www.clarin.eu/resource-families).
[6] Perseus Digital Library (accesible desde: http://www.perseustufts.edu/hopper), Scaife Viewer (accesible desde: https://scaife.perseus.org), Laboratoire d'Analyse Statistique des Langues Classiques (accesible desde: https://web.philo.ulg.ac.be/lasla).
[7] GALICIANA (accesible desde: https://biblioteca.galiciana.gal).
[8] Chadwyck-Healey Literary Collections (accesible desde: http://collections.chadwyck.com).
[9] Biblioteca Virtual Miguel de Cervantes (accesible desde: http://www.cervantesvirtual.com).

textos libremente accesibles es muy limitada, las operaciones de numerización (escaneo y oceri-zación) son fastidiosas y caras, los criterios de constitución de un corpus son exigentes[10]. Y cuando los corpus no están construidos con criterios sólidos y claros, ligados a unos objetivos de investi-gación muy reflexionados, la investigación se cae: no basta el cuanto más, mejor.

Lo cierto es que el tratamiento de los datos textuales que podemos captar y los cambios en la escala de observación requieren herramientas digitales que nos permitan la observación densa y fina de las prácticas culturales y de sus actores culturales, más aún la formulación de otras pregun-tas o las de siempre, pero sobre corpus mucho más amplios: necesitamos una ayuda para la clasifi-cación, la descripción, la comparación, la publicación, la reescritura. Y son muchas las operaciones que podemos realizar con textualidades electrónicas para su estudio: desde el enriquecimiento de textos para la edición didáctica o científica, hasta la pura re-creación artística para libros electrónicos o arte digital. Pero aquí nos hemos centrado en las herramientas y metodologías digi-tales que permiten el análisis cuantitativo y cualitativo de textos.

Estos análisis son herederos de la lingüística empírica que siempre ha buscado describir usos lingüísticos atestiguados por su número, contrariamente a otras corrientes y métodos que con cierta sorna los más empíricos llaman *lingüística de sillón*: aquella basada en los ejemplos que se le ocur-ren al lingüista desde ¡su sillón!

Esta aproximación se alió con las posibilidades del cálculo desde el siglo XIX en psico-lingüística, a principios del siglo XX para las distribuciones léxicas, con la estadística lingüística en los años 60: es la lingüística del corpus, la lingüística matemática. De ahí, fácilmente en nuestro siglo hemos llegado al tratamiento de lenguas naturales, la minería de datos o la inteligencia artificial, para desarrollar ejercicios estilométricos, reconocimiento de temas, etnografías digitales, etc. Se trata de un acercamiento probabilista al texto que es tratado como un saco de palabras del que se extraen muestras y cuyas distancias se miden respecto a las medias.

El análisis cuantitativo se interesa por las ocurrencias (el número de apariciones) y concor-dancias, mide correlaciones, busca relaciones necesarias. Contamos con herramientas y funcionali-dades básicas en abierto que pueden ser utilizadas en diferentes etapas de la formación secunda-ria y superior como el pequeño Docuburst, el inmenso N-Grams Viewer, Voyant Tools o Sketch En-gine11, o bien otras ya consagradas en diferentes tradiciones como Wordsmith, Hyperbase, Ant-Conc, etc.[12]. De la familiarización con estas herramientas por parte de los estudiantes e investiga-dores a la necesidad de hacer su propia programación con Python o con R para responder a pre-guntas de investigación suyas, hay un paso que la formación reglada también tiene que dar.

---

[10] Ioana Galleron et al. (2018). Las publicaciones digitales de corpus de autores. Guía de trabajo, plantilal de análisis y recomendaciones (trad. de Paloma Ortega Deballon) (accesible desde: https://halshs.archives-ouvertes.fr/halshs-02164065).

[11] Docuburst (accesible desde: http://vialab.science.uoit.ca/portfolio/docuburst), N-Grams Viewer (accesible desde: https://books.google.com/ngrams), Voyant Tools (accesible desde: https://voyant-tools.org/), Sketch Engine (accessible desde: https://www.sketchengine.eu/).

[12] Wordsmith (accesible desde: https://wordsmith.org/), Hyperbase (accesible desde: http://hyperbase.unice.fr/hyperbase/), AntConc (accesible desde: https://www.laurenceanthony.net/software/antconc/).

El análisis cualitativo estudia un fenómeno en contexto –las apariciones de una(s) expresión (es) en un conjunto documental– porque se interesa por el contenido semántico del enunciado o sus papeles pragmáticos. Se trata de una investigación basada en anotaciones, sean automáticas con analizadores automáticos morfológicos o sintácticos, esto es, con una codificación cerrada y estandarizada; sean manuales, de acuerdo con una metodología y un modelo de anotaciones semántico o pragmático, esto es, con una codificación abierta, de forma que el investigador o lector toma contacto con el material y va elaborando hipótesis y etiquetas que luego se articulan y se seleccionan. Por anotación entendemos un texto breve que viene a enriquecer el texto inicial (soporte) en un lugar preciso (por ejemplo, una palabra) que se convierte así en marcado (mark-up).

Se trata, pues, de una operación de explicitación y de recodificación del texto por cuanto que se crean datos (etiquetas) y desde ahí se elaboran teorías explicativas de un fenómeno particular. Se apoya en prácticas intelectuales de reagrupamiento, de inducciones generalizadoras o incluso de transposiciones. Es lo que podemos hacer por ejemplo con CATMA, ATLAS.Ti, o NVIVO[13], unas capaces de anotar solo texto, pero otras también el texto y la imagen y el sonido y el vídeo.

En todos los casos, es una forma de subjetividad asumida por cuanto que controlada metodológicamente y permite cierta reproductibilidad del análisis, aunque sea difícil compartir un etiquetado guiado por una pregunta particular y un modelo interpretativo.

Señalemos, para terminar, que ambos tipos de análisis pueden utilizar métodos inductivos cuando no hay hipótesis antes de examinar datos y resultados, como la *corpus-driven linguistics* o *usage-based linguistics* (guiada por el corpus), y/o métodos deductivos a partir de una hipótesis ya elaborada (un término, una funcionalidad, un valor), a la manera de la *corpus-based linguistics* (fundada en el corpus).

Este monográfico está destinado a mostrar los primeros resultados de investigaciones con estas herramientas y metodologías, pero hemos de pensar en las posibilidades didácticas que nos ofrecen en los diferentes grados de la educación secundaria y superior. Y ello porque permiten: (1) un recorrido organizado desde un problema o pregunta a la constitución de un corpus, a la elección de una herramienta en función de los datos que esa herramienta necesita y de los resultados que puede ofrecer, pasando por una determinación del nivel de análisis en el que se va a desarrollar la observación; (2) esa observación puede ser sistemática y completa a niveles diferentes, en contextos muy reducidos o muy amplios; (3) la atención puede centrarse en las palabras que aparecen. Esto es: permiten leer despacio, tomar conciencia de la (meta)lectura, dialogar. O si lo preferimos: permiten encontrar e interpretar, contar e interpretar, visualizar e interpretar.

Pero saltan los problemas: la dificultad y el tiempo requeridos para constituir los corpus tanto en el ámbito sociológico (las entrevistas, los tuits, los flujos de comentarios de todo tipo), como en el espacio de los estudios literarios; las limitaciones de todas estas herramientas que ofrecen tales funcionalidades y tales no, por no hablar de sus precios que no están al alcance ni del estudiante, ni del investigador, ni de los grupos, apenas de las instituciones (y sus responsables) que no

[13] CATMA (accesible desde: https://catma.de/), ATLAS.Ti (accesible desde: https://atlasti.com/), NVIVO (accesible desde: https://nvivo-spain.com/)

las conocen aún. Finalmente, hay tensiones, pues la ciencia busca regularidades, normas y leyes, pero encuentra hápax: "La heroica ciudad dormía la siesta" es única, su estructura es repetida.

Estos artículos ofrecen aproximaciones, primero, con herramientas cuantitativas, luego cualitativas, unas en el marco de la lingüística y de la etnografía digital, las otras en el espacio de los estudios literarios.

Primero, Elena del Olmo e Iván Arias proponen, en "Un estudio empírico con Sketch Engine sobre la interfaz sintáctico-pragmática para la identificación de la estructura temática en español", una utilización casi exhaustiva de la herramienta Sketch Engine en busca de marcadores morfosintácticos o léxicos de los roles de tema y rema sobre un corpus en español de textos periodísticos, y ello a partir de la teoría de la progresión temática. Estos patrones han de mostrar la secuencia de ideas o conceptos que se desarrollan a lo largo de un texto, una representación de enorme interés en todas las áreas de la textualidad en Humanidades y Ciencias Sociales. Los investigadores pasan, a la hora de descubrir posibles patrones relevantes, del análisis automático del conjunto al análisis manual de las primeras apariciones, para volver a comprobar si los patrones de aparición observados en la anotación manual se constatan en la totalidad del corpus. Las limitaciones provienen del enfoque eminentemente léxico de la herramienta y, de ahí, la necesidad de buscar instrumentos de búsqueda con una mayor capacidad expresiva en cuanto a los patrones de dependencia sintáctica.

Inés Pérez Fresno, en "Las herramientas digitales en el análisis de traducciones: una aproximación cualitativa al análisis de la traducción de literatura infantil", también utiliza Sketch Engine para su análisis comparativo de la novela para niños *Les malheurs de Sophie* de la Condesa de Ségur (1858) y dos de sus traducciones al español: *Las desventuras de Sofía* (1970) y *Las desdichas de Sophie* (2018), pero esta vez para confirmar los primeros resultados obtenidos con Stylo y Gephi[14], en una triangulación metodológica que le permite reivindicar la rapidez en la obtención de resultados, la posibilidad de tratar grandes volúmenes de texto, así como la facilidad de uso de estas herramientas que proporcionan al investigador una base inicial de trabajo para cualquier investigación cualitativa posterior.

Después, Héctor Puente, Costán Sequeiros y Marta Fernández, en "Discursos sociales en *Cyberpunk 2077*: un estudio de caso de los debates sociopolíticos de la comunidad de videojugadores en Youtube", desarrollan un estudio del discurso y comentarios sobre *Cyberpunk 2077*, un videojuego de rol de mundo abierto que ha sido muy prolífico en la emergencia de debates y activismo en comunidades online como YouTube. A partir de una conceptualización que proviene de Roger Caillois, Michel Foucault, Pierre Bourdieu y Erving Goffman, proponen un diseño inductivo y artesanal, más intensivo que extensivo, para comprender las significaciones otorgadas por los usuarios a las experiencias y prácticas sociales presentes en el juego, usando herramientas cualitativas de corte etnográfico digital como es ATLAS.ti. Practican estos autores una codificación híbrida entre la generación automática de códigos de categorización y un refinado manual que permiten

[14] Stylo (accesible desde: https://github.com/computationalstylistics/stylo), GEPHI (accesible desde: https://gephi.org/).

la codificación, vinculación y análisis de los diversos comentarios recogidos. No se trata simplemente del procesamiento de flujos de datos masivos o del potencial análisis a tiempo real, sino de la posibilidad de acceder a las estructuras y relaciones subyacentes entre los usuarios/as, a la viralización y extensión de sus producciones discursivas y a la representación mediante grafos de redes sociales que se establecen entre ellos.

Finalmente, Adrián Menéndez de la Cuesta, en "Modelo de análisis cualitativo con Atlas.ti para novelas postdigitales: *Reina* y *Game Boy*", aborda el estudio de *Game Boy* de Víctor Parkas (2019) y *Reina* de Elizabeth Duval (2020) desde un marco conceptual que llama postdigital y nos sitúa en este momento del siglo. Propone un enfoque hermenéutico que requiere herramientas de análisis cualitativo como ATLAS.ti. Para ello, explica cómo la fase de construcción del modelo interpretativo resulta crucial: si las anotaciones realizadas con la herramienta sobre las obras aportan los datos que se recuperarán e interpretarán en la fase de análisis, el modelo traduce las preguntas de investigación en códigos para la herramienta. Ese proceso para definir los códigos siempre es deductivo-inductivo y el enfoque hermenéutico logra detectar matices, alusiones veladas a lo postdigital, incluso omisiones significativas que resultan indetectables para herramientas cuantitativas. Así demuestra que lo contemporáneo postdigital no alude a un periodo diferenciado, sino a un solapamiento de rupturas y permanencias según diferentes ritmos, clave de las obras analizadas.

En todos los artículos observamos una permanente ida y vuelta y regreso entre polos que no son opuestos sino los recorridos: permanentemente pasan del papel al formato electrónico en la pantalla y de nuevo al papel, del análisis cuantitativo al cualitativo, de una lectura desde lejos (masiva) a una lectura pegada al texto (único): de lo que se repite al hápax. Más aún, pasan de una aproximación heurística a otra hermenéutica y de nuevo heurística. También vemos iterar la perspectiva inductiva capaz de hacer visibles construcciones que de otra manera no vemos, a partir de palabras clave (las más frecuentes); y la perspectiva deductiva para verificar la validez de una construcción que conocemos de antemano, confirmar una hipótesis, construir un sentido.

A estas trayectorias de lecturas que se cruzan permanentemente con el uso de herramientas y metodologías digitales, queremos llamarlas lecturas re-mediadas, lecturas remediadas, quién sabe si remediadoras.