FACULTAD DE HUMANIDADES Y CIENCIAS DE LA EDUCACIÓN

Grado en Filología Hispánica

UNIVERSIDAD DE JAÉN
Facultad de Humanidades y Ciencias de la Educación

Trabajo Fin de Grado

Filología y Humanidades Digitales: un análisis de "Deshoras" de Cortázar desde la teoría de los grafos.

Alumno/a: **Mariia Chizhikova**

Tutor/a: Prof. D. David González Ramírez
Dpto.: Filología española

Mayo, 2022

1. INTRODUCCIÓN 2

2. HACIA LAS HUMANIDADES DIGITALES 5

2.1. La Cuarta Revolución Industrial 5

2.2. De archivo a base de datos 6

2.3. Análisis computacional de textos literarios.....7

2.4. Humanidades Digitales como práctica generativa 10

2.5. Edición digital vs Digitalización 12

3. MARCO PRÁCTICO: DEL GRAFO AL TEXTO, DEL TEXTO AL GRAFO.....18

3.1. Conceptos básicos: la teoría de grafos 18

3.2. La teoría de grafos en la investigación literaria.....22

3.3. Metodología 24

3.3.1. Definición del conjunto de vértices del grafo.....24

3.3.2. Construcción del grafo 26

3.3.3. Comentario del texto 27

4. CONCLUSIONES 34

5. BIBLIOGRAFÍA 37

1. INTRODUCCIÓN

La democratización de los medios informáticos hizo que el porcentaje de población mundial que usa Internet aumente de 29% a 60% ente 2010 y 2020, según el Banco Mundial[1]. El fenómeno global de la progresiva digitalización y el consecuente aumento continuo de la interconexión social retan y transforman las rutinas y prácticas existentes en todas las esferas de la vida. Semánticamente, sería más apropiado denominar 'burbujas' lo que acabamos de llamar 'esferas', dado que las fronteras entre los distintos ámbitos tienden a difuminarse, pues cada vez más se habla del saber unitario. La gente en todos los rincones del planeta participa activamente en el consumo y la creación del contenido web que puede ser infinitamente más variado en cuanto a su forma que cualquier fuente material de información. Al mismo tiempo surgen las nuevas tendencias en el arte que aprovechan lo digital como una extensión de la libertad creadora.

Los entornos virtuales, lejos de ofrecer meramente un espacio alternativo para la comunicación humana, transforman la percepción y ofrecen una flexibilidad sin precedente para la interacción, la creación y la investigación. Al mismo tiempo, el intento de abordar la naturaleza cambiante del conocimiento actual plantea desafíos metodológicos para la ciencia marcando la insuficiencia de la estricta disciplinariedad. El mencionado desafío metodológico consiste en la necesidad de disponer de conocimientos de distintas disciplinas para enfrentarse a la investigación.

El presente trabajo es fruto de curiosidad que nació a raíz del auge de la inteligencia artificial (IA). Los ingenieros que se enfrentan al reto de dotar las máquinas de habilidades propias del ser humano (comprender y generar el lenguaje, analizar imágenes y sonido, tomar decisiones) hace mucho tiempo que reúnen equipos de varios perfiles, porque dichas tareas requieren conocimiento de una variedad impresionante de ámbitos: desde la ciencia cognitiva hasta la ingeniería informática. Por otra parte, muchas disciplinas aprovechan los avances de

[1] https://datos.bancomundial.org/indicator/IT.NET.USER.ZS

la IA para progresar, de modo que, por citar un ejemplo, surgen estudios de marketing en base de aplicación de los algoritmos de aprendizaje automático para modelar el comportamiento de los consumidores.

Mi primera aproximación a la fascinante sinergia que surge en el cruce de la IA y la informática con las disciplinas humanísticas acaeció en el curso 2020-2021, cuando hice el trabajo que me llevó a dar conferencia titulada «Las interfaces de voz: hacia una naturalidad artificial» en el seminario del grupo de investigación ALTYA en abril de 2021. La indagación sobre el funcionamiento de las interfaces de voz me hizo consciente de que el progreso exige colaboraciones más inesperadas, como la de los especialistas en fonética con los informáticos y estadísticos y plantó en mi mente la semilla del deseo de ir más allá de los límites de la estricta disciplinariedad.

Desde abril de 2021 llevé a cabo varios trabajos en los que ponía a prueba mis recién adquiridos conocimientos informáticos en conjunto con mi formación filológica. En noviembre de 2021 se publicó mi estudio sobre clasificación automática de artículos científicos relacionados al COVID-19 según su tema (Chizhikova, 2021). En diciembre del mismo año, en contexto de un trabajo académico, recopilé y publiqué en acceso abierto un corpus de 18,172 reseñas de hoteles andaluces etiquetados según la polaridad de los sentimientos que expresan[2]. Luego sobre esta colección de textos desarrollé una red neuronal capaz de clasificar reseñas según su polaridad con 0.96 de precisión[3].

Estos trabajos, sin embargo, no implicaban la faceta literaria de mi formación y se situaban más en el área del Procesamiento de Lenguaje Natural (PLN) y la Lingüística computacional, mientras que todos los trabajos literarios que realizaba y leía se quedaban en el terreno de la filología tradicional. Viendo muy de cerca la manera en la que la metodología computacional impulsa el desarrollo de disciplinas al parecer muy ajenas a la informática, surgió la idea de indagar en la faceta interdisciplinar de las Humanidades.

Las Humanidades Digitales, al aplicar los enfoques computacionales a la investigación humanística, han expandido significativamente el horizonte de alcance de la investigación cultural. Enfatizando en la creación, conservación y interpretación del código cultural de la

[2] Disponible en: https://www.kaggle.com/datasets/chizhikchi/andalusian-hotels-reviews-unbalanced

3 Disponible en: https://www.kaggle.com/code/chizhikchi/lstm-binary-sentiment-classification-for-spanish#Introduction

humanidad, las Humanidades Digitales aprovechan las ventajas de los entornos digitales para abrir nuevas perspectivas, plantearse nuevas preguntas y crear nuevos recursos de producción e interpretación de la experiencia humana.

Con el objetivo de ofrecer una perspectiva interdisciplinar y polifacética hacia una disciplina de fronteras fluidas y cambiantes, presentamos este trabajo que analiza el fenómeno de las Humanidades Digitales centrándose en su dimensión filológica. La primera parte sienta una base teórica a través de una revisión histórica del la transformación social que acaece actualmente y del impacto que ha tenido la confluencia entre las disciplinas humanísticas y las ciencias de la computación. Se indagará en las particularidades del *modus operandi* de los filólogos digitales de hoy en día para intentar responder a una serie de cuestiones metodológicas y prácticas sobre la edición digital de textos y mostrar cómo la revolución digital se manifiesta en la filología moderna.

En la segunda parte del trabajo se presenta un nuevo método de análisis literario con el fin de explorar la potencialidad de la exploración textual cuantitativa. Se pretende, por tanto, trastocar la dicotomía entre lo cuantitativo y lo cualitativo poniendo las teorías y las herramientas de lo que tradicionalmente se conocía como *ciencias exactas* a disposición de la filología. En concreto, nos ocuparemos de la conversión del cuento *Deshoras* de Julio Cortazar en un grafo de conocimiento para construir un fundamento objetivo para la caracterización de la obra literaria.

2. HACIA LAS HUMANIDADES DIGITALES

2.1. La Cuarta Revolución Industrial

El término *La cuarta revolución industrial* fue acuñado por Klaus Schwab, el fundador y el presidente ejecutivo del *Foro económico mundial*, que ya en 2015 afirmaba que la revolución tecnológica, cuyo proceso empezó a mediados del siglo pasado, iba a cambiar drásticamente nuestras vidas en un abrir cerrar los ojos: «We stand on the brink of a technological revolution that will fundamentally alter the way we live, work, and relate to one another. In its scale, scope, and complexity, the transformation will be unlike anything humankind has experienced before» (Schwab, 2015: s.p.).

La palabra *revolución* denota un 'cambio rápido y profundo en cualquier cosa'[4]. Estos cambios radicales se han producido a lo largo de la historia de la humanidad: el paso de la recolección a la agricultura hace aproximadamente diez mil años; la revolución agraria; la Primera revolución industrial (s. XVIII) debida a la invención de las máquinas de vapor; la Segunda revolución industrial (ss. XIX-XX) impulsada por al advenimiento de la electricidad; la Tercera revolución industrial provocada por la proliferación de los ordenadores y el Internet a finales del siglo pasado. La Cuarta revolución industrial que estamos presenciando estriba en la Tercera y, como señala Schwab (2016: 12) es caracterizada no solo por el desarrollo de los sensores y procesadores más potentes y el fomento de la inteligencia artificial, sino por la manera en la cual estos avances interactúan con los distintos dominios, tales como el psíquico, digital y biológico:

> The fourth industrial revolution, however, is not only about smart and connected machines and systems. Its scope is much wider. Occurring simultaneously are waves of further breakthroughs in areas ranging from gene sequencing to nanotechnology, from renewables to quantum computing. It is the fusion of these technologies and their interaction across the physical, digital and biological domains that make the fourth industrial revolution fundamentally different from previous revolutions (Schwab, 2016: 12)

[4] DLE (2014), *s.v.* «revolución»

Como vemos, precisamente esta integración casi instantánea de los avances tecnológicos recientes en todas las disciplinas científicas (incluso las que tradicionalmente se consideraban muy lejanas de la tecnología) y en la vida cotidiana de cada persona hace que la Cuarta revolución industrial sea especial en la historia de la humanidad. Gracias a esta revolución hoy en día hablamos de tales disciplinas como las Humanidades Digitales, el PLN o la Lexicografía digital.

Una vez hecha esta revisión, aunque muy rápida, de la transformación social, tecnológica y científica que todavía está en curso, podemos pasar a ver más detalladamente cómo este proceso venía influyendo sobre él las Humanidades llevando esta disciplina al ámbito tecnológico.

2.2. De archivo a base de datos

El primer contacto entre la humanidades y las ciencias de la computación se produjo a mediados del siglo pasado con una serie de proyectos archivísticos. De este modo, el primer paso hacia el estado actual de la disciplina venía marcado por una necesidad de estructuración de los datos para facilitar el acceso a ellos. Una vez agotado el potencial del almacenamiento físico del saber y en vista del inmenso potencial que translucían los avances de la computación, se optó por iniciar un proceso de digitalización de las fuentes materiales: «these activities typically focused on corpus building, on creating standarts for text enconding, and on building databases that could facilitate work on humanistic corpora» (Burdick *et al.* 2012: 8). La digitalización se ha convertido en un requisito indispensable para la estructuración de los datos tras la Cuarta Revolución Industrial, pero su grado no siempre ha sido lo bastante alto y varía de un proyecto a otro: una colección de documentos en formato .pdf, como, por ejemplo, la *Biblioteca digital hispánica*, es, desde luego, una colección estructurada, indexada por los metadatos de cada obra (título, autor, fecha de publicación y etc.). Sin embargo, no está estructurada a nivel de los textos de las obras, cuyas ediciones no son ediciones digitales, sino meras digitalizaciones. Nos ocuparemos más adelante de las diferencias que hay entre estos dos conceptos y de la importancia que tiene la estructuración de los datos, ya que en este apartado el foco de nuestra atención recae sobre la trayectoria histórica del desarrollo de la disciplina.

Entre los años 80 y el inicio del s. XXI se planificaron y fueron realizados varios proyectos que Burdick *et al.* (2012: 8) agrupa bajo el la etiqueta de primera oleada de Humanidades Digitales (eng. *first wave of Digital Humanities*). Uno de los proyectos más ambiciosos de esta época es el proyecto de la *Biblioteca digital Perseo* (eng. *Perseus Digital Library Project*) que se ideó en 1985 con el fin de explorar el fenómeno la digitalización de las bibliotecas y su paso al Internet: «Perseus is a practical experiment in which we explore possibilities and challenges of digital collections in a networked world» (Smith, et al. 2000). Cabe señalar, que las mismas ideas subyacen al proyecto de la ya citada *Biblioteca digital hispánica* que se presentó, según Corral (2009), en enero de 2008 con 10.000 documentos y hoy en día, según la *web oficial*, incorpora mas de 222.000 títulos de todas las temáticas y tipos documentales.

2.3. Análisis computacional de textos literarios

Hasta ahora veníamos citando los proyectos cuyo principal objetivo fue la creación, difusión y conservación de los recursos digitalizados, lo cual constituye solamente una rama de las Humanidades Digitales. Ya en el s.XX, como una continuación directa de la transición del saber al terreno digital, surgió el ámbito del análisis computacional del texto, cuyo fundador fue Roberto Busa, quien tras defender su tesis doctoral en 1946 empezó a idear el *Index Thomisticus* (un corpus de concordancias de las palabras en las obras de Santo Tomás de Aquino). Este proyecto fue fundamental para la consolidación de todas las disciplinas que combinan el saber humanístico y el de las ciencias exactas, por lo cual conviene detenernos en las aportaciones más importantes de Busa.

Index Thomisticus es el fruto de más de treinta años de trabajo en colaboración con el gigante computacional IBM. Aunque el resultado final consistiera en la indexación de 10,631,980 palabras en formato de tarjetas perforadas, el proyecto tuvo muchas facetas: fue uno de los primeros intentos de computerización del lenguaje que impulsó la educación de la gente en el ámbito (Busa organizó una escuela de operadores de tarjetas perforadas) y de creación de los diccionarios automáticos (eng. *machine dictionary*) como el *Lexicon Electronicum Latinum* compilado para lematizar las palabras del *Index*. Además, puso de relieve la importancia de colaboración entre las humanidades y las ciencias de la computación y sentó bases para los proyectos en una disciplina que en tiempos de Busa todavía carecía de nombre.

Al presentar el *Index Thomisticus* Busa remonta a los años cuarenta, el momento de la escritura de su tesis, cuando sobre dos conclusiones preliminares se sentaron las bases de su *opus magnum*. En primer lugar, se dio cuenta de la insuficiencia de la lectura para el entendimiento del sistema conceptual subyacente al texto, por lo cual se hacían necesarias aproximaciones de otro tipo: «a philological and lexicographical inquiry into the verbal system of an author has to precede and prepare for a doctrinal interpretation of his work» (Busa, 1980: 83). Esto se debe a que para una correcta interpretación del sistema conceptual que un autor expresa a través de su sistema verbal hace falta desprenderse de los sesgos que impone el sistema conceptual del lector. El análisis cuantitativo del lenguaje es, por su naturaleza, una técnica de mayor objetividad que proporciona datos sobre los cuales se puede construir un nuevo punto de vista hacia el sistema conceptual del autor de una obra.

En segundo lugar, recalcó la importancia de las palabras funcionales, muchas veces denominadas 'vacías de significado': «I realized that all functional or grammatical words (which in my mind are not 'empty' at all but philosophically rich) manifest the deepest logic of being which generates the basic structures of human discourse» (Busa, 1980: 83). Esta premisa se encuentra en la base de otra rama de las Humanidades Digitales que se puede considerar una aplicación específica del análisis computacional del texto: la estilometría. Aunque la primera aproximación cuantitativa al análisis del estilo se remonta al año 1901, cuando T. Mendenhall presenta su método basado en el cómputo de números de letras en cada una de las palabras de las obras de Shakespeare, se puede afirmar que Busa adelanta las consideraciones fundacionales de la estilometría moderna, disciplina que pone énfasis en el análisis de uso de las palabras funcionales. Desde el principio de los estudios estilométricos se intentaba encontrar un método que permitiera resolver el problema de la atribución de autoría en base a los métodos cuantitativos. El primer gran alcance la disciplina se produjo años después de la publicación de la tesis de Busa, en 1963 con el famoso caso de los *Federalist papers*, una colección de 85 artículos y ensayos escritos por Alexander Hamiltom, James Madison y John Jay bajo el seudónimo colectivo Publius. La autoría de doce de aquellos artículos fue objeto de un largo debate que impulsaba a los investigadores a inventar nuevos métodos y enfoques para conocer si fueron escritos por Hamilton o Madison.

Mosteller y Wallace en 1963 estudiaron la cuestión aplicando técnicas estadísticas de análisis de datos basadas en el teorema de Bayes. Las variables seleccionadas para la discriminación fueron las frecuencias de uso de las palabras funcionales. El empleo de las

funciones discriminatorias y métodos Bayesianos permitió a los estudiosos llegar a la conclusión de que Madison es el autor más probable de los artículos discutidos. Asimismo señalaron que las palabras gramaticales pueden ser una fuente muy valiosa de la información para resolver problemas de autoría, probando así la afirmación de Busa: «The function words of the language appear to be a fertile source of discriminators and luckily the hight-frequency words are the strongest» (Mosteller, Wallace, 1963: 306).

Además de los indicios sobre los orígenes de los métodos empleados hoy en día, los primeros trabajos en el ámbito fronterizo entre las humanidades y ciencias demuestran la conciencia y entusiasmo de los autores de estar abriendo un nuevo camino del desarrollo, de estar pisando un terreno sin explotar: Monsteller y Wallace hicieron observaciones sobre el teorema de Bayes recalcando la carencia de estudios sobre la variación de resultados con distintos antecedentes de un método que en el año 1963 cumplió dos siglos. Busa tuvo que persuadir a los expertos del líder el sector de las tecnologías de la información y comunicación (TIC) de su época de que su proyecto no era imposible. Los pioneros planteaban el problema teórico y metodológico del uso de los ordenadores y la aplicación de métodos cuantitativos en datos humanísticos.

Roberto Busa no solo desarrolló un método que se sigue utilizando para la investigación lingüística y literaria (sirvió de modelo para el Corpus Diacrónico del Español de la RAE, por ejemplo), sino también adelantó muchos de los cambios que experimentaría la disciplina con la Cuarta Revolución Industrial y formuló una serie de principios que luego se consolidaron como fundamentos no solo de las Humanidades Digitales, pero también de la Lingüística Computacional, la Lexicografía Digital y el PLN. Al final de su artículo, reflexiona sobre el papel del ordenador en las Humanidades y llega a la conclusión de que las tecnologías no tienen que utilizarse únicamente para acelerar el proceso de la investigación. Por el contrario, predice que el uso de las ciencias de la computación en la investigación humanística tiene que alargar los proyectos y aumentar el trabajo de los científicos que ahora se ocuparán de análisis en un nivel superior, cumpliendo con las funciones creativas, más propias y características de la mente humana: «the use of computers in the humanities has as its principal aim the enhancement of the quality, depth and extension of research» (Busa, 1980: 89). En los apartados que siguen veremos qué impacto tuvo esta nueva distribución de tareas en los proyectos que se llevan a cabo hoy en día.

2.4. Humanidades Digitales como práctica generativa

La predicción de Roberto Busa de que los ordenadores iban liberar a los investigadores de trabajo repetitivo de pre-procesamiento de datos empezó a cumplirse incluso antes de que se acabara su proyecto, a medida de que se iban mejorando los ordenadores. Con lo arraigado que están las TIC en nuestro día a día, puede resultar sorprendente que el equipo de Busa procesase 95 millones de palabras en 25 años y el mismo investigador, ya en el año 1980 evidencia el rápido desarrollo que tuvieron tales tecnologías, como el reconocimiento óptico de caracteres (OCR por sus siglas en inglés), afirmando que en su momento tuvieron que resolver problemas que dejaron de existir en la fecha de la publicación del artículo: «I imagine that a similar quantitiy of text could be processed today in ten or even six years, using video terminals and optical character recognition» (Busa, 1980: 87). Actualmente, los sistemas de OCR como Convertio tardan aproximadamente diez segundos en procesar una página en formato A4, cifra que puede disminuir si se dispone de mejores recursos computacionales[5].

Además, se han automatizado y, consecuentemente, acelerado tales cómputos, como el de número de letras en cada palabra (empleado por Mendenhall en 1901) o el de las frecuencias de aparición de los términos (usado por Monsteller y Wallace en 1963). Esta automatización abrió el camino a nuevos métodos de extracción de características del texto, como la frecuencia del término-frecuencia inversa (eng. *term frequency-inverse document frequency*) (Salton *et al,* 1988) o los vectores semánticos de palabras (eng. *word embeddings*) (Grave *et al*, 2018). Los enfoques del análisis textual también experimentaron una transformación de modo que de la aplicación de simples teoremas y procedimientos estadísticos se pasó a diseñar y aplicar algoritmos de aprendizaje automático y, posteriormente, redes neuronales que actualmente se consideran sistemas de estado de arte para abarcar numerosas tareas, tales como la clasificación de textos, el reconocimiento de entidades nombradas, traducción automática, generación automática de resúmenes, etc.

Sin embargo, la tecnología no tiene que considerarse exclusivamente como el instrumento de investigación, sino también como su objeto, porque las Humanidades Digitales van mucho más allá de la estructuración y análisis de datos textuales. La Cuarta revolución industrial hizo que nos comuniquemos de una forma mucho más variada y amplia que el texto lineal. La información se convirtió en el oro de nuestra generación y su adecuada exploración

[5] https://support.convertio.co/hc/en-us

determina el desarrollo de la sociedad actual. Estos cambios, evidentemente, también conciernen a los datos humanísticos y deben moldear la metodología de la disciplina que se ocupa de su estudio. Esto, según Burdick *et al.* (2012: 10), implica la extensión del concepto de las letras hacia disciplinas tales como el diseño gráfico (eng. *graphic design*), narrativa visual (eng. *visual narrative*) y el desarrollo de interfaces de usuario. A la lista deben añadirse también la ciencia de datos y su sub-área de PLN. Consecuentemente, en el paso hacia las Humanidades Digitales acaece la reinterpretación de la disciplina como una práctica generativa, una práctica de crear cosas (programas, visualizaciones, sitios web y bases de datos) y no solo conocimiento.

La extensión del ámbito humanístico hacia una gran variedad de disciplinas y prácticas emergentes produce un panorama heterogéneo cuyas fronteras son difíciles de delimitar. Esta característica de la disciplina se ve reflejada en su denominación: el sintagma 'Humanidades Digitales' no es solamente un sustantivo colectivo que hace referencia a un conjunto de disciplinas, sino también designa a las Humanidades en plural: ciencias capaces de abarcar y relacionar objetos divergentes de naturaleza lingüística, literaria, histórica, mediática, tecnológica, etc.

Rojas Castro, en un estudio sobre los principios, valores y prácticas de las Humanidades Digitales, señala que «el núcleo de las Humanidades Digitales corresponde al proceso de digitalización» (Rojas Castro, 2013: 85), dado que, según él, todo empleo de herramientas computacionales implica traducir un signo analógico al signo digital. Creemos que este razonamiento no toma en consideración el grado de arraigue que tienen las TIC en la sociedad actual. Vivimos en un mundo en el que los hallazgos neurocognitivos se utilizan como fundamentos del diseño intuitivo de las interfaces gráficas del usuario (Meneses *et al.*, 2021) y se trabaja sobre creación de interfaces de voz entre muchas otras vías de fomento de la integración lo más fluida posible con las herramientas tecnológicas. Por lo tanto, afirmar que el primer objetivo de las Humanidades Digitales es convertir lo análogo en lo digital significa subestimar el grado de desarrollo de la tecnología y de la disciplina, dado que hoy en día el advenimiento de los datos masivos ha hecho que a menudo se requiera una traducción de lo digital a lo análogo y no al revés.

Un ejemplo muy ilustrativo de este fenómeno es el modelado del comportamiento del usuario a partir de los registros del navegador web y del buscador (Qin, Z. *et al*., 2020), una

tarea que requiere creación un sistema multimodal capaz de relacionar el texto de la búsqueda con los datos de la actividad del usuario, datos cuya naturaleza es puramente digital.

Las convocatorias de congresos, como el *V congreso de la Sociedad Internacional de Humanidades Digitales Hispánicas* (HDH 2021), o de artículos en revistas, como la *Revista de Humanidades Digitales* (RHD), podrían ayudar a definir la disciplina, pero en vez de marcar sus fronteras las difuminan incorporando temas y prácticas que van del análisis de discurso hasta representaciones 3D, realidad virtual y realidad aumentada[6]. Esto se debe al hecho de que la digitalización del conocimiento y su incesante movimiento por las redes requiere y a la vez posibilita la creación de nuevas vías de creación y manejo de la información de distinta índole. La disciplina encargada de estudio de la cultura de la era digital, al igual que su objeto de investigación, tiene que prescindir de límites e huir de la heterogeneidad en el intento de establecer vínculos hasta ahora impensables entre ámbitos divergentes: «however heterogeneous, the Digital Humanities is unified by its emphasis on making, connecting, interpreting, and collaborating» (Burdick *et al*., 2012: 24).

2.5. Edición digital vs Digitalización

La computación depende de la desambiguación en todos los niveles y se basa en procedimientos explícitamente definidos que se ejecutan paso a paso. Estos principios, al parecer, contradicen la metodología humanística, porque la ambigüedad y lo implícito desempeñan un papel fundamental en las Humanidades. Es por ello que en el cruce entre las Humanidades y las ciencias de la computación se tienen que buscar compromisos y maneras de conjugar enfoques tan divergentes para encontrar maneras de llevar lo implícito y lo ambiguo al terreno computacional.

Los libros en papel y las ciencias humanísticas comparten sus historias en el sentido de que los humanistas han trabajado con este formato durante siglos. El paso del papel a lo digital implica, como ya habíamos notado, muchos cambios que requieren adaptar la metodología humanística al lenguaje computacional a la vez que abren nuevos caminos de exploración textual. Las operaciones cuantitativas, tales como la construcción de concordancias, el cálculo de frecuencias de términos o la indexación de tokens actualmente se situarían en el nivel básico del procesamiento textual y sirven para responder rápidamente a

[6] Llamamiento para contribuciones del HDH 2021

algunas preguntas sobre los datos textuales. El formato digital agiliza muchos procesos, hace contraer nuevas rutinas y manejar herramientas cada vez más complejas.

Por otro lado, las metodologías de análisis más sofisticadas requieren mayor estructuración de datos para abrir la dimensión interpretativa en el flujo digitalizado y cuantificado de caracteres del teclado. Según señala Calvo Tello (2019: 116), «el diseño y la recogida de datos define en buena medida los pasos o las conclusiones a las que podemos llegar», por lo cual la construcción de proyectos de Humanidades Digitales depende del conocimiento de los fundamentos computacionales de la metodología que se quiere aplicar.

El concepto de la edición digital surge de la necesidad de estructuración de datos textuales y es muy importante establecer una frontera rígida entre lo *digital* y lo *digitalizado*. En términos básicos, una edición *digital* no puede imprimirse sin que pierda una parte considerable de su funcionalidad, que se consigue, por ejemplo, mediante enriquecimiento con las etiquetas de lenguaje de marcado XML-TEI de los datos textuales obtenidos en el proceso de *digitalización* (Fradejas Rueda, 2009-2010). Estas etiquetas permiten marcar una parte del texto para que luego se pueda extraer, manejar y explorar aplicando los métodos computacionales.

Es difícil poner en duda la importancia de una buena preparación de los datos para su futura exploración y sin embrago, según Calvo Tello (2019: 116), «hay una laguna de responsabilidad sobre quién debe asumir la ingente tarea de seleccionar, filtrar y poner a disposición datos que el resto de la comunidad de investigadores puedan utilizar de manera completa».

Budrick *et al.* (2012: 17-18) esbozan la estructura de los proyectos de Humanidades Digitales en varios niveles, situando las actividades de curación (eng. *curation*), análisis, edición y modelado en un nivel inmediatamente superior al de las actividades estrictamente computacionales. Estos procesos se califican como fundamentales, pues constituyen el núcleo de la disciplina y pueden ayudar a definir mejor el concepto de la edición digital.

Se entiende por *curación* la selección y organización de materiales en una estructura interpretable. Los medios digitales dan mucho espacio a la creatividad de los Humanistas Digitales a la hora de plantear la estructura de datos y la manera de presentarlos. Un correcto tratamiento de textos determina el valor e impacto que puede tener el proyecto en su totalidad, dado que el advenimiento de Big Data hizo patente la necesidad de tener colecciones de datos

fiables, sostenibles y lo máximo universales posible. Por la misma razón la ecdótica está viviendo una transformación importante para continuar siendo una actividad imprescindible: «the parsing of the cultural record in terms of questions of authenticity, origin, transmission, or production is one of the foundation stones of humanistic scholarship upon which all other interpretive work depends» (Budrick *et al.*, 2021: 18). El concepto de modelado se refiere a la búsqueda de una manera adecuada de presentación las los datos. En este estadio un humanista digital se encargaría de transformar el conocimiento generado en un conjunto de contenidos culturales organizados según una o varias características distintivas, «this a project dedicated to analyzing the correspondence of a famous artist might assume a chronological shape, which is one model of human life» (Budrick *et al.*, 2021: 19).

La diversidad de posibilidades de presentación de la información en el dominio digital hace que en el proceso de la edición digital se tengan que tomar decisiones importantes en relación con el formato de los archivos de almacenamiento, el tamaño del corpus y sus metadatos. El objetivo del proceso de la edición siempre es polivalente, en tanto que concierne al objeto de investigación (¿Qué textos queremos recopilar y analizar?), a la metodología (¿Qué técnica deseamos aplicar?) y al resultado final que se quiere alcanzar (recordemos que las Humanidades Digitales son una práctica generativa).

Según Calvo Tello (2019: 117), «el tipo de análisis determina en cierta medida tanto las épocas como los géneros [literarios] que podemos investigar». Algunos de los métodos de análisis ampliamente usados por expertos en distintas áreas de conocimiento, tales como el aprendizaje automático (ML por sus siglas en ingles) y aprendizaje profundo (DL por sus siglas en inglés), requieren grandes cantidades de datos para entrenar modelos y en el caso del ML y DL supervisado hace falta que estos datos sean anotados. Aunque actualmente se esté trabajando en mejorar el rendimiento de los modelos de IA en escenarios en los que no se dispone o se dispone de muy pocos datos anotados (*zero-shot* o *few-shot learning*), los experimentos se centran en otros dominios, como el clínico (Sivarajkumar *et al*, 2022), por ejemplo.

Dependiendo del tipo de edición que se quiere generar y del modelo de su presentación al público, se tiene que adoptar un formato maestro, el «formato principal del proyecto que es controlado manualmente y que, en caso de encontrarse errores o de quererse ampliar la información, puede ser modificado» (Calvo Tello, 2019: 124). A continuación se ofrece una

revisión de los rasgos característicos de algunos de los formatos más frecuentemente elegidos actualmente para realizar una edición digital o digitalizada.

El formato PDF es el utilizado mayoritariamente para las ediciones digitalizadas, porque es fácil de generar y no requiere un esfuerzo de lectura y manejo por parte de un usuario acostumbrado a manejar fuentes en papel. Sin embargo, las posibilidades de estructurar y analizar el texto almacenado en este formato son muy limitadas o requieren su transformación a otro formato mediante procedimientos como el OCR. Por lo tanto, si se aspira a una edición verdaderamente digital de una obra o un corpus textual, conviene hacer uso de otros formatos.

El entorno digital ofrece herramientas para aglutinar varias versiones de un mismo texto, seguir su desarrollo, las variantes que presenta. La integración de las nuevas herramientas es idónea para el desarrollo de ediciones críticas: «the use of structured and/or tagged approaches to identify persons, themes, places, or features of a text provides a way to maximize the intellectual investigation of documents and to display these interpretations» (Budrick *et al.*, 2021: 35). Uno de los métodos para la realización de una edición digital que más libertad ofrece es el formato XML-TEI que en realidad constituye un estándar de representación de las características en recursos textuales que tienen que ser identificados explícitamente para facilitar su procesamiento por otros programas (TEI Consortium, 2022).

XML (*Extensible Markup Language*), como indica su nombre, es un lenguaje de marcado caracterizado por ser extensible, es decir que el conjunto de etiquetas utilizadas lo puede definir cualquier editor según sus requisitos. El estándar TEI especifica un conjunto de *labels* que pueden ser manejados en un intento de facilitar un intercambio sostenible de datos entre los usuarios. Al mismo tiempo, sirve como base para el diseño de sistemas y creación de nuevos programas para la exploración textual. Calvo Tello (2019: 126) señala entre las desventajas de este formato «el hecho de que esté menos extendido que HTML», que «sea más complicado de aprender» y que le «falte soporte en muchas herramientas». No obstante, este formato está cada vez más extendido entre la comunidad científica. Según Google Scholar, desde el año 2021 hasta la mayo de 2022 se publicaron aproximadamente 2900 artículos académicos relativos al uso del marcado XML-TEI.

Otra opción de almacenamiento de colecciones de textos son las bases de datos relacionales, que ofrecen la ventaja de integración fluida con herramientas de desarrollo web, pero su manejo normalmente requiere conocimientos de SQL (*Sructured Query Language*), su

funcionamiento eficiente demanda una estructuración muy meticulosa de los datos y no facilitan la exploración de datos a nivel de texto.

El formato más simple y universal, pero también el más semánticamente sencillo, es el texto plano. Los documentos .txt carecen de formato y se usan «principalmente cuando el texto ha de ser entregado a algún programa que debe analizarlo o anotarlo» (Calvo Tello, 2019:125).

Como hemos señalado en el principio de este apartado, el formato maestro tiene que ajustarse a los objetivos que cada uno de los proyectos, si bien es cierto que se deben tener en cuenta la pervivencia del recurso lingüístico-literario en la web para su reutilización en estudios posteriores. Para hacerlo posible, se desarrollan estándares, como el TEI que, sin limitar considerablemente la libertad que puede ofrecer el entorno digital a un autor, sirven de guía para creación de productos compatibles con otras herramientas.

En relación con esto, conviene subrayar la existencia de varios tipos de anotación textual, entre los cuales destaca el marcado lingüístico. La anotación lingüística de un texto se puede llevar a cabo en los niveles fonético, morfosintáctico y semántico, siendo el segundo el más extendido debido a que es útil para estudios estilométricos y a que categorías morfosintácticas de las palabras pueden usarse como características adicionales para el entrenamiento de redes neuronales. Otro tipo de anotación se situaría en el nivel de la crítica textual para registrar, por ejemplo, las tachaduras, las adiciones, las enmiendas, las erratas o las variantes en el proceso de desarrollo de una edición crítica digital.

Además de ofrecer un texto (estructurado con anotaciones o no) en un formato legible tanto para los usuarios como para las máquinas, una edición digital tiene que ofrecer una serie de metadatos, al igual que cada libro tiene su portada y cada documento del archivo (su ficha). Los metadatos imprescindibles para cada obra editada digitalmente son título, autor y fecha de publicación, pero esta lista se puede extender hasta abarcar toda la información recopilada sobre el objeto de estudio. Estas etiquetas poseen no solo carácter informativo, sino también ayudan a estructurar las entradas de un corpus para hacer posibles tales operaciones como búsqueda de obras por autor, fecha, etc. Los metadatos desempeñan un papel fundamental para el manejo de grandes colecciones permitiendo plantear nuevas cuestiones sobre las relaciones que mantienen las obras entre sí, con el contexto histórico, con la geografía, la economía, etc.

La publicación de las ediciones digitales en la web se hace mayoritariamente bajo las licencias *Creative Commons* que permiten, de manera sencilla y estandarizada dar permiso a compartir y usar los recursos digitales bajo condiciones que defina el autor. Dado que la ciencia se basa en la aportación de los hallazgos y descubrimientos precedentes, la publicación de resultados de la investigación en acceso abierto contribuye al progreso global. Además, la publicación de los datos de manera libre «revierte en la sociedad la inversión pública en investigación, posibilita la ciencia replicable, refuerza las Humanidades Digitales en español» (Calvo Tello, 2019: 131).

La edición digital es una de las modalidades que puede tener un proyecto de Humanidades Digitales, cuya elaboración consiste en una constante adaptación de los principios de las Humanidades tradicionales a los nuevos entornos digitales, siendo la variabilidad su principal ventaja. El polimorfismo y la flexibilidad de lo digital ofrece muchas posibilidades para hacer realidad trabajos que nunca podrían ser llevados a cabo en papel, pero a la vez plantea un reto de homogeneización de formatos para aprovechar al máximo los proyectos realizados haciéndolos compatibles con unos con otros. Consecuentemente, resulta imposible e innecesario dar una definición exacta del concepto de la edición digital «ya que sus infinitas posibles funcionalidades y recursos hacen de ella un concepto vago e indeterminado» (Mateos Frühbeck, 2019: 240). No obstante, en el ejemplo del diseño de los recursos textuales digitales se hace aún más patente la redefinición de las disciplinas humanísticas en prácticas generativas que permite aumentar su horizonte de análisis hacia casi el infinito.

3. MARCO PRÁCTICO: DEL GRAFO AL TEXTO, DEL TEXTO AL GRAFO

De acuerdo con los principios básicos de las Humanidades Digitales desarrollados en la parte teórica del presente trabajo, el principal objetivo de la parte práctica consiste en presentar un nuevo método para el análisis literario encaminado a la construcción de un fundamento objetivo y cuantitativo para la investigación humanística. En concreto, analizaremos un cuento de Julio Cortázar, *Deshoras* (1982), recogido en el homónimo volumen de cuentos, mediante la construcción de un grafo de conocimiento que vinculará los conceptos relevantes del texto y los personajes. La extracción de los conceptos, los personajes y sus relaciones se llevará a cabo mediante un programa que desarrollamos en el lenguaje *Python* mientras que para el almacenamiento y la visualización de los datos se usará la plataforma *Neo4j*[7]. De este modo, primero daremos cuenta del método para extracción del grafo de conocimiento y luego construiremos un análisis de *Deshoras* desde el punto de vista de las Humanidades Digitales, apoyándonos en el análisis del grafo.

3.1. Conceptos básicos: la teoría de grafos

Antes de describir el método propuesto, tenemos que dar cuenta de algunos de los conceptos básicos de la TEORÍA DE GRAFOS con los que vamos a operar en este trabajo. Leonhard Euler, matemático y físico suizo, es considerado el fundador de la llamada TEORÍA DE GRAFOS, que es una rama de Matemática Discreta cuyo objetivo es explorar «la conectividad entre los elementos de un conjunto» para resolver «cierto tipo de problemas que únicamente tienen que ver con un determinado número de puntos y ciertos trazos que los unen» (Menéndez Velázquez, 1998: 11). En su trabajo *Solutio problematis ad geometriam situs* (Euler, 1741) presentó ante la Academia de San Petersburgo la solución del famoso problema de los puentes de Koenigsberg. El matemático reconoció que Leibnitz fue el primero en fijarse en la existencia unos principios que existen «Praeter illam Geometriae

[7] https://neo4j.com

partem, quae circa quantitates versatur, et omni tempore summo studio est exculta»[8] (Euler, 1741: 128). Estos principios fueron englobados en una disciplina titulada *Geometriam situs*, aunque Euler enfatiza sobre la ausencia tanto de una tipología de problemas pertenecientes al ámbito de la disciplina, como de una metodología propia.

El problema de los puentes de Koenigsberg consistía en plantearse si se puede pasar por todos los siete puentes que unen una isla en el río Pregolia (fig. 1) con el resto de la ciudad sin pasar por cada uno más de una vez. A partir de este problema concreto, Euler formuló un problema más general: «quaecunque sit flucii figura et distributio in ramos, atque quicunque fuerit numerus pontium, inuenire, utrum per singulos pontes semel tantum transiri queat, an vero secus?»[9] (Euler, 1741: 129).

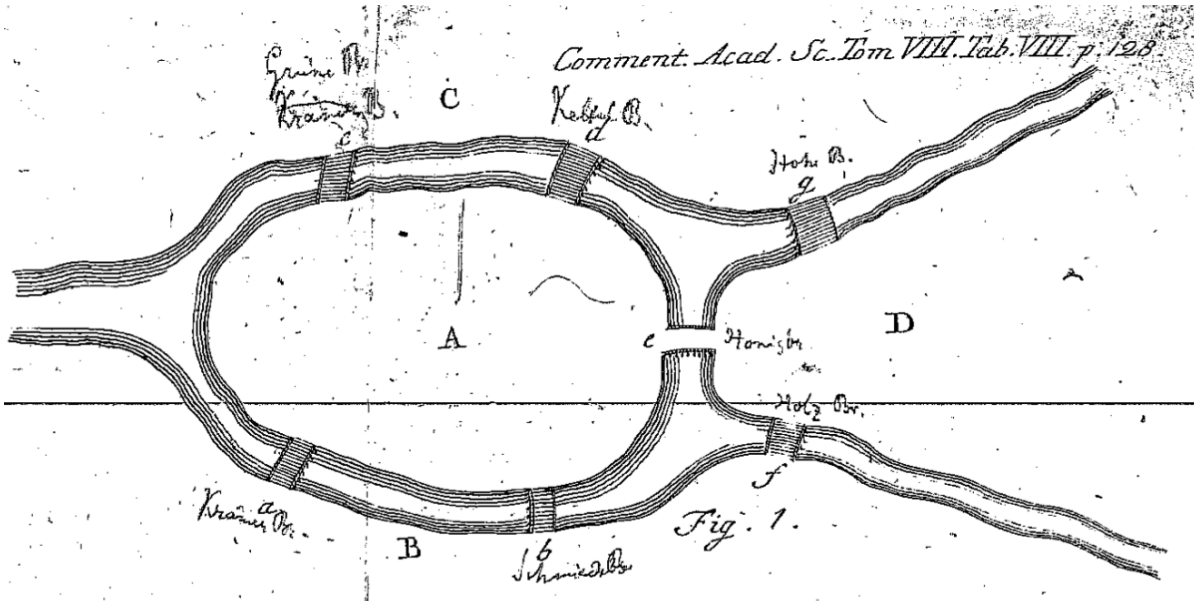

Figura 1. Esquema de los puentes de Koenigsberg (Euler, 1741 :128)

Para encontrar la solución, como se ve en el esquema, Euler denominó a cada región con una letra mayúscula. El científico argumento que el camino por todos los puentes sin pasar por ninguno dos veces sería posible si el número de conexiones de cada región es par. Un requisito adicional sería que el punto inicial y final coincidan. Por lo cual, no se puede pasar

[8] esp.: «Además de esa parte de la geometría, que gira en torno a las cantidades, y que ha sido cultivada en todas las épocas con el mayor celo» (traducción propia)

[9] esp.: «Cualquiera que sea la forma del río, y su distribución en los brazos, y cualquiera que sea el número de puentes que haya, averigüe si se puede cruzar solo una vez por cada puente, o de otra manera» (traducción propia)

por los siete puentes de Koenigsberg sin repetir ningún recorrido, porque las cuatro regiones tienen un número impar de conexiones (B, C y D tienen tres, A - cinco).

No obstante la destrucción de los siete puentes originales por el bombardeo de Koenigsberg (1944) durante la Segunda Guerra Mundial, el problema resuelto por Euler sigue siendo relevante para la actualidad, pues, como ya mencionamos anteriormente, dio pie a la noción de grafo, que es un tipo de estructura de datos utilizada ampliamente en ciencias de la computación. La teoría de grafos tiene múltiples aplicaciones posibles:

> se utilizó para la resolución de juegos matemáticos, para el estudio de circuitos eléctricos y en diversas aplicaciones en una multitud de campos tan diferentes como la economía, física teórica, psicología, física nuclear, lingüística, sociología, zoología, tecnología, antropología, computación, biología, ingeniería, química (Menéndez Velázquez, 1998: 11).

El grafo consta, geométricamente, de un conjunto de VÉRTICES o NODOS V y de un conjunto de ARCOS o ARISTAS A que se representa con el par G = (V, A) = (V(G), A(G)). El número de vértices de un grafo se denomina el ORDEN DEL GRAFO y el conjunto de sus aristas determina el TAMAÑO DEL GRAFO.

Un grafo puede simbolizar múltiples sistemas que contienen elementos relacionados entre sí, como por ejemplo, los temas, las ideas y los conceptos dentro de una obra literaria. Hay que tener en cuenta, sin embargo, que «un grafo contiene únicamente información topológica, es decir, información sobre la conectividad entre puntos, careciendo de información geométrica en el sentido euclídeo (distancias, ángulos...)» (Menéndez Velázquez, 1998: 12). Por lo tanto, es importante poder conocer los caminos que recorren las aristas de un grafo. Un CAMINO o ruta de un grafo G es una secuencia finita en la que aparecen los vértices V y aristas A «donde cada lado tiene por extremos los vértices inmediatamente precedente y siguiente en la secuencia» (Menéndez Velázquez, 1998: 13). Los caminos pueden ser ABIERTOS o CERRADOS (dependiendo de si el punto inicial y final coinciden). Un camino cerrado se denomina CIRCUITO si todos los nodos menos el primero y el último son distintos. La DISTANCIA $d(x, y)$ de x a y es la longitud del CAMINO más corto de x a y o ∞ si el camino no existe. Si en un grafo G cada par de vértices x, y está conectado, el grafo se denomina CONEXO. Es más, un grafo NO CONEXO, puede contener un SUBGRAFO FUERTEMENTE CONEXO.

Una distinción muy importante que conviene apuntar el la que se da entre los GRAFOS DIRIGIDOS y GRAFOS GENERALES. «Un dígrafo (o GRAFO DIRIGIDO) es una clase especial de

grafos, en los cuales a las aristas se les asigna un sentido, el cual se representa geométricamente mediante un flecha» (Menéndez Velázquez, 1998: 16).

Además de la distancia, otro concepto muy importante para nuestro estudio es la CENTRALIDAD DE NODO, cuya determinación es un objeto de polémicas dentro de la teoría de grafos. «Idealmente cada nodo [...] ejerce un grado de influencia dentro de su determinado dominio y se espera que esta influencia pueda alumbrar la estructura del grafo; la centralidad es una medida cuantitativa que pretende revelar la importancia de un nodo» (Boldi, 2014: 223). No existe una única fórmula para medir la centralidad de un nodo, sino que se tiene que elegir un enfoque en función de cada estudio en concreto: «the important question in considering applications -and the one that is most often neglected- involves considering the relevance of the particular structural attribute measured to the substantive problem being studied» (Freeman, 1977: 39).

Boldi y Vigna en su artículo de 2014 proponen una serie de axiomas de centralidad (eng. *axioms of centrality*) en las que recogen una serie de propiedades básicas y simples que tiene que mostrar la medida de centralidad. Con este enfoque axiomático consiguieron demostrar que la fórmula más representativa es la CENTRALIDAD HARMÓNICA. La centralidad harmónica es una medida geométrica, es decir, se basa en la suposición de que la importancia de un vértice depende de las distancias que le separan de otros. La centralidad armónica soluciona el problema que presentaba la noción de *cercanía* a la hora de trabajar con grafos que presentan nodos inalcanzables $d(x,y) = \infty$.

Sin embargo, para nuestro trabajo sería interesante también trabajar con la CENTRALIDAD DE INTERMEDIACIÓN que es una medida basada en caminos, es decir, explora no solo la existencia del camino más corto, sino que examina todos los caminos más cortos que llevan al nodo. Dicho en otras palabras, mayor importancia tiene un nodo cuya eliminación daría lugar a conversión de más vértices en terminales. Las medidas de centralidad basadas en la intermediación fueron introducidas por Freeman: «the importance of this conception of point centrality is in the potential of a point for control of information flow in the network. Positions are viewed as structurally central to the degree that they stand between others and can therefore facilitate, impede or bias the transmission of messages» (Freeman, 1977: 34). En el desarrollo de nuestro trabajo argumentaremos cuál de las dos medidas (cercanía harmónica o centralidad de intermediación) se adecúa mejor a los objetivos de la presente investigación.

3.2. La teoría de grafos en la investigación literaria

El término GRAFO DE CONOCIMIENTO fue popularizado después del año 2012, cuando fue introducido *Google´s Knowledge Graph* (Singhal, 2012) que transformó la página de los resultados del famoso buscador Google añadiendo información sobre enlaces relacionados, biografía de las personas, sus obras, etc. Se señala que la creación del grafo de conocimiento de Google es un gran paso hacia la construcción de un buscador de nueva generación que aprovecha la inteligencia colectiva de sus usuarios y entiende el mundo de una manera más parecida a la de los humanos (Singhal, 2012). En este ejemplo vemos que los grafos de conocimiento pueden ser muy generales y cubrir grandes cantidades de dominios específicos y servir para mejora de la experiencia de un número inmenso de los usuarios.

Haslhofer *et al*. (2018: 2) sitúan la construcción de los grafos de conocimiento para Humanidades Digitales dentro del conjunto de sistemas de organización del conocimiento (eng. *knowledge organization systems*). De este modo, la construcción de un grafo constituye una metodología de estructuración del conocimiento que permite enfatizar en las relaciones entre los datos. Conviene resaltar que esta metodología no viene a sustituir a las otras, sino sirve de extensión a los enfoques cualitativos y cuantitativos de análisis. A continuación se ofrece una revisión histórica de la paulatina adaptación que experimentó esta metodología al utilizarse para la investigación humanística.

La creación y exploración de los grafos de conocimiento fue adaptada por los sociólogos para analizar las relaciones dentro de las comunidades que forman las personas. El descubrimiento de la estructura dinámica subyacente a la sociedad que determina todas sus agrupaciones periféricas y formales dio lugar a consolidación de una disciplina titulada SOCIOMETRÍA por el psicólogo estadounidense Jacob Levy Moreno en 1941. Su objetivo principal consistió en investigar las relaciones de atracción y repulsión entre personas y como estas se proyectan en redes psicosociales a gran escala. Levy Moreno definió conceptos como *átomo social* (eng. *social atom*) -el patrón de proyecciones de atracciones y repulsiones entre los miembros de un grupo- y *red psicosocial* (eng. *Psycho-Social Network*) -conjunto de individuos que, sin importar la lejanía geográfica o social, asociados unos con otros por medio de vínculos y contra-vínculos de atracción-repulsión, pueden compartir entre sí noticias, opiniones, etc. (Moreno, 1941: 24-27). Moreno fue el primero en aplicar la teoría de los grafos o redes (eng. *network theory*) a la descripción de los fenómenos sociales.

La introducción de los métodos cuantitativos en los estudios culturales se denominó CULTURONOMÍA (eng. *culturonomics*) en el trabajo de Michel y Lieberman Aiden, *et.al* (2010). Los investigadores analizaron un corpus de 5,195,769 libros (4% de todos los libros publicados hasta la fecha) y sacaron conclusiones sobre la evolución del lexicón y gramática ingleses, sobre las tendencias a evocar el pasado, la pervivencia de las tendencias, el impacto de las censuras y otras limitaciones. También analizaron tales fenómenos culturales, como el feminismo, el impacto de guerras y otros conflictos, de descubrimientos científicos, de la religión e incluso se pudo estudiar la evolución de la dieta estadounidense.

En cuanto a los estudios literarios, la teoría de los grafos junto con los postulados sociométricos abrió el paso a la exploración computacional de las redes sociales en las obras literarias. El género que más se estudio con estos métodos es, sin duda, el teatro. Esto se debe a la relativa facilidad con la que se extraen las relaciones entre los personajes de una obra teatral tan convenientemente, desde el punto de vista computacional, dividida en escenas y dispuesta en un formato fijo, en el sentido de que no cabe ambigüedad de atribución de las réplicas a los personajes. Todo esto, junto con la relativa brevedad de las obras teatrales, propició la creación de tales corpus anotados en XML-TEI, como *DraCor*[10] (Fisher, 2019).

Este proyecto explora el concepto de un *corpora programable* (eng. *programmable corpora*) que además de proporcionar el acceso a los datos mediante una API (interfaz de programación de aplicaciones), se construye como una infraestructura digital para la investigación literaria que contiene herramientas para la visualización de las redes sociales entre otras. El proyecto *DraCor*, como podemos ver, encaja a la perfección con los principios de las Humanidades Digitales que formulamos en los apartados anteriores, porque no solo implica la actividad de creación del conocimiento, sino también se ocupa del diseño de herramientas y nuevos métodos.

La colección a día de hoy consta de trece subcorpus en distintas lenguas, entre los cuales encontramos dos colecciones en castellano: *Calderón drama corpus*[11] y *Spanish Drama Corpus*[12]. El primero consistente de 201 dramas y fue preparado por un grupo de especialistas del Instituto de Lenguas y Literaturas Romances de la Universidad de Tübingen y el Instituto

[10] https://dracor.org

[11] https://dracor.org/cal

[12] https://dracor.org/span

de Estudios Romances de la Universidad de Viena. El segundo constituye una adaptación a la plataforma *DraCor* del corpus BETTE (Biblioteca Textual del Teatro en Español de 1868-1934) (Calvo Tello *et al.* 2015).

Estas colecciones favorecieron la exploración de las redes sociales en teatro, que desembocaron en numerosos trabajos en los que se analizan múltiples obras para describir la evolución de géneros teatrales (Wendell, 2021), de su poética (Ehrlicher *et al.*, 2020), de las acotaciones (Maximova y Fisher, 2019) o de la complejidad estructural de las obras a lo largo del tiempo (Fisher, *et al.*, 2017).

El éxito de la plataforma DraCor es muy significativo para nuestro trabajo, porque es, desde luego, un proyecto-modelo de Humanidades Digitales que marca las pautas para los compiladores de corpus digitales. Por otro lado, el hecho de que haya dado lugar a numerosas y muy diversas investigaciones subraya la importancia de anotación de obras literarias (un proceso manual largo y complejo que sin embargo se compensa por la apertura de nuevos caminos para la ciencia, como es la aplicación de una teoría proveniente de matemática discreta a la filología).

No podemos desatender el hecho de que en todas las investigaciones revisadas la teoría de los grafos se aplica al análisis de las redes sociales, a la relación entre personas reales o personajes de las obras literarias la mayoría de las cuales son teatrales. En cambio, en el presente trabajo se propone aplicar dicha teoría a obras menos estructuradas externamente, a los cuentos, y trabajar no solo con los personajes, sino también con los conceptos o ideas.

3.3. Metodología

3.3.1. Definición del conjunto de vértices del grafo

La extracción automática de los elementos relevantes de un texto y vínculos entre ellos es un reto que va más allá de las Humanidades Digitales al constituir una de las tareas sobre las que se trabaja en PLN. Para esta tarea actualmente se emplea una gran variedad de técnicas, entre las cuales sobresale el uso de las redes neuronales, algoritmos estadísticos que calculan las probabilidades que tiene cada una de las palabras del texto de constituir un elemento de interés o no. Sin embargo, para entrenar una red neuronal se necesitan enormes cantidades de datos previamente etiquetados de las que no disponemos, porque no existe ningún corpus de cuentos etiquetado a nivel de palabra con marcas de conceptos relevantes.

Una vez desechada la posibilidad de basarnos en una red neuronal para la extracción de ideas relevantes, hizo falta definir el objeto de nuestra búsqueda. El *Diccionario de la lengua española* explica la palabra *concepto* como 'idea que concibe o forma el entendimiento' en la primera acepción y como 'representación mental asociada a un significante lingüístico' en la sexta[13]. De estas dos definiciones llegamos a establecer el objetivo de buscar las reglas a las cuales se ajusta la distribución de las palabras más significativas en los cuentos de Cortázar.

En su ensayo *De cuento breve y sus alrededores* (1969) Cortázar, basándose en la doctrina de Horacio Quiroga, reflexiona sobre los rasgos que influyen en la calidad de un cuento evocando el proceso creativo desde la elección del tema hasta la escritura, puesto que según él «la idea de significación no puede tener sentido si no la relacionamos con las de intensidad y de tensión que ya no se refieren solamente al tema sino al tratamiento literario de ese tema». Es precisamente la definición de la *intensidad* lo que nos arroja luz sobre la adecuada metodología de tratamiento de cuentos de Cortázar, porque se afirma que «consiste en la eliminación de todas las ideas o situaciones intermedias, de todos los rellenos o fases de transición que la novela permite e incluso exige» (Cortázar, 1969). De acuerdo con este razonamiento, un buen cuento debe prescindir de todas ideas que no sean esenciales, es decir, no tiene que incorporar ninguna palabra irrelevante, dado que «la única forma en que puede conseguirse este secuestro momentáneo del lector es mediante [...] un estilo en el que los elementos formales y expresivos se ajusten, sin la menor concesión, a la índole del tema» (Cortázar, 1969). Así pues, para extraer los significantes lingüísticos de todos los elementos relevantes del cuento hace falta identificar todos los sustantivos del texto, suponiendo que dos elementos están vinculados si co-ocurren en un espacio de quince palabras dentro del texto.

Más allá de la selección de una correcta metodología de extracción de términos relevantes del cuento, el ensayo citado nos permite argumentar la pertinencia de la aplicación de la metodología sugerida a las obras de Cortázar. El autor, al reflexionar sobre la elección de un buen tema para su cuento llega a la conclusión de que este, paradójicamente, no debe «de ser extraordinario, fuera de lo común, misterioso o insólito», sino al contrario «puede tratarse de una anécdota perfectamente trivial y cotidiana» (Cortázar, 1969). De especial relevancia para este trabajo resulta la comparación que establece el cuentista entre un buen tema y un sistema atómico, es decir un núcleo en torno al cual giran los electrones, y la afirmación de que «un bien tema atrae todo un sistema de relaciones conexas» (Cortázar, 1969). Si el tema

[13] DLE, 2021, *s.v.* concepto

del cuento construye o activa en el autor y posteriormente en el lector una serie de nociones, el texto del cuento tiene que reflejarlas, lo que hace posible su extracción y análisis de las relaciones entre ellas.

De este modo, entre los elementos que formarán parte del conjunto de los vértices del grafo de conocimiento extraído del texto de *Deshoras* figurarán nodos de dos tipos: personaje y concepto. Entre los personajes tendremos en cuenta a las entidades humanas que se identifican por su nombre (no entramos en esta parte en reflexiones sobre el desdoblamiento de la personalidad del narrador): Sara, Doro, Aníbal y Felisa. La lista de los conceptos relevantes se definirá mediante la extracción automática del texto de todos los sustantivos. Para esta identificación se usará un modelo de análisis lingüístico de la biblioteca SpaCy[14] que incorpora un mecanismo de etiquetado de las palabras con sus clases morfológicas.

3.3.2.Construcción del grafo

Una vez definido el conjunto de términos que formarán parte de los vértices del grafo, tenemos que proceder con la extracción de las relaciones entre ellos. Como adelantamos en el párrafo anterior, suponemos que dos elementos están vinculados entre sí si se encuentran a distancia igual o superior a quince palabras. En este apartado describiremos el procedimiento de creación del grafo de conocimiento mediante tratamiento computacional del texto de *Deshoras* convertido en formato de texto plano y guardado en un archivo .txt.[15]

Como hemos mencionado anteriormente, nos servimos de la biblioteca SpaCy para automatizar el análisis lingüístico. SpaCy proporciona a sus usurarios los llamados *pipelines*, es decir, conjuntos de tecnologías involucradas en el proceso de procesamiento de datos textuales dividido en fases por las que el texto pasa igual que pasa el agua por una tubería, con la diferencia de que los datos de salida del *pipeline* se modifican de modo que salen más estructurados.

Para nuestro proyecto optamos por un *pipeline* basado en el modelo *es_core_news_sm* que fue entrenado sobre el corpus *AnCora* (Taulé, M. *et al*, 2008). Esta herramienta es formada de varios componentes, tales como un transformador de tokens en palabras (*tok2vec*), un etiquetador morfológico, un *parser* de dependencias sintácticas, un lematizador y un

[14] https://spacy.io

[15] El código disponible en https://github.com/chizhikchi/oddhours

sistema de reconocimiento de entidades nombradas. La aplicación de este análisis al texto del cuento nos permite localizar todos los sustantivos léxicos en el texto que llegan a formar un conjunto de 267 vértices de tipo «Concepto» del futuro grafo de conocimiento. El conjunto de nombres de los personajes fue definido manualmente: Sara, Doro, Aníbal y Felisa.

Una vez extraídas las entidades que formarán parte de los dos conjuntos de los vértices del grafo, podemos pasar a extraer relaciones entre ellos según la regla que definimos anteriormente: suponemos que dos entidades están vinculadas si se encuentran a distancia de menos de quince palabras dentro del texto. Para inferir las relaciones existentes entre los conceptos y los personajes, recurrimos a un matcheo basado en reglas (*rule-based matching*) usando la clase *Matcher* de SpaCy. Conviene notar que también se guardó la información sobre la frecuencia de relaciones entre los elementos como propiedad de cada una de las aristas.

Una vez extraídos los datos necesarios para la creación del grafo de conocimiento, lo guardamos en una base de datos remota *Neo4j Sandbox*. *Neo4j* es una plataforma de datos especializada en grafos que proporciona distintas soluciones para sus usuarios. Para llevar a cabo la presente investigación optamos por una base de datos temporal y gratuita, *Neo4j Sandbox*, que está optimizada para análisis, modelado y visualización de datos.

Para almacenar nuestros datos en la nube, utilizamos la extensión de *Neo4j* para el lenguaje Python que permite interactuar con la base de datos sin cambiar el entorno de trabajo. Por último, usamos la herramienta *NEuler* para calcular las métricas de centralidad de intermediación y centralidad harmónica de cada uno de los nodos y extraer el subgrafo fuertemente conexo. Estadísticas sobre los nodos con los mayores valores de ambas méticas se pueden ver en las figuras tres y cuatro. Una vez generado el grafo de conocimiento a partir del texto de *Deshoras*, podemos proceder con el comentario del cuento.

3.3.3.Comentario del texto

En este apartado se ofrece una revisión de algunos de los aspectos fundamentales del cuento cortazariano, tales como su estructura, tiempo interno o técnica narrativa. Con este análisis se intentará aprovechar la caracterización del grafo de conocimiento como metodología de exploración cuantitativa que, combinada con la exploración cualitativa tradicional, permita abrir una nueva perspectiva crítica propia de las Humanidades Digitales.

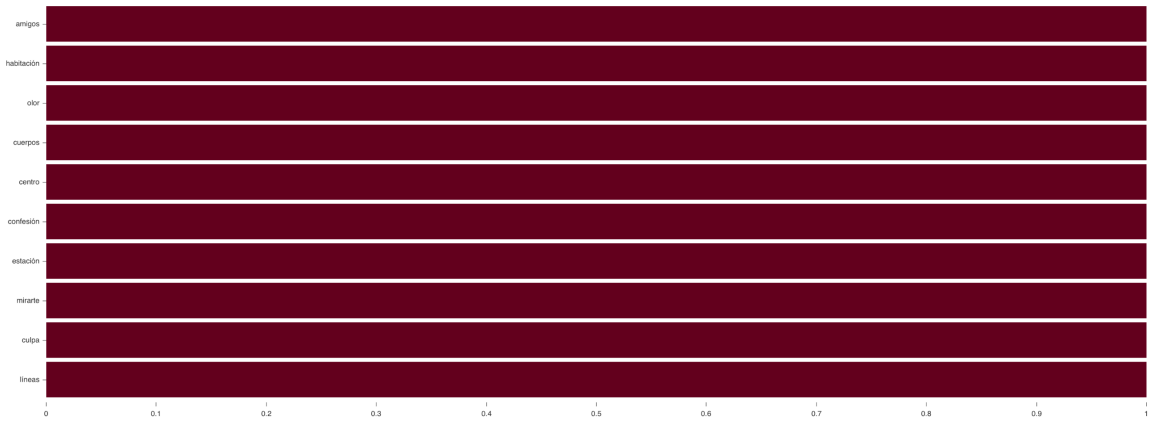

Figura 3. 10 nodos con mayor valor de centralidad harmónica

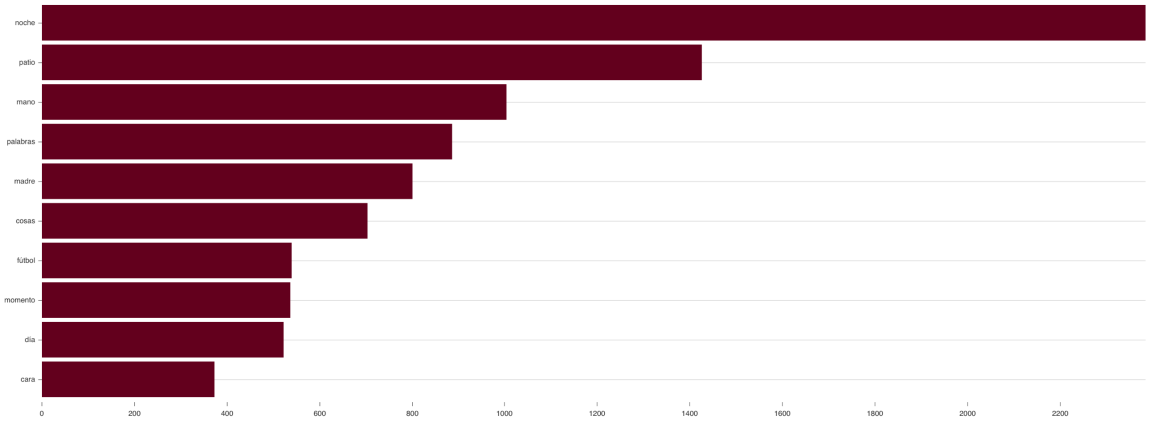

Figura 2. 10 nodos con mayor valor de centralidad de intermediación

Deshoras (1982) es el octavo y el último volumen de cuentos publicado por Julio Cortázar. El hecho de que el título del cuento analizado fue elegido para representar una colección de ocho historias nos lleva a prestarle atención al paratexto que hace referencia, según afirma el autor en una de sus entrevistas, a «una no coincidencia en el tiempo, destinos que pasan uno al lado del otro sin encontrarse, sin juntarse» (Cortázar, 1983), es decir, al tema principal del texto comentado. En el diccionario de la RAE se recogen dos definiciones de la locución adverbial *a deshora*: 'fuera de sazón o de tiempo' y 'de repente, intempestivamente'[16] y ambas son significativas para el presente trabajo. Para ofrecer una visión más amplia sobre la estructura del cuento que se erige en el manejo del tiempo *a deshora*, primero ofrecemos un breve resumen de su argumento.

El relato trata sobre un hombre que se sumerge en los recuerdos de su infancia en Bánfield, de los viejos tiempos en los que pasaba todos los días con su mejor amigo Doro. De

[16] RAE (2014) *Diccionario de la lengua española*, *s.v* 'deshora'

una manera muy especial recuerda a Sara, la hermana mayor de su antiguo mejor amigo, su primer amor infantil. Al final del cuento, la escritura del protagonista se interrumpe con una llamada a la cotidianidad que corta la ilusión de los recuerdos.

Conviene notar que desde el principio de la historia se hace patente que el narrador es consciente de que los recuerdos de su infancia no tienen cabida en su cotidianidad. Es más, emergen en su vida, como emerge lo fantástico en la literatura, provocando una ruptura, un choque entre los dos mundos. Este hecho hace cuestionarnos acerca del género del cuento que a simple vista no muestra ninguna característica que permitiría situarlo dentro de lo fantástico dado que aparentemente se trata de un hombre que se pone a escribir sus recuerdos en un arrebato de nostalgia. Sin embargo, en su ensayo *El sentimiento de lo fantástico*, Cortázar niega la obligatoria excepcionalidad de lo fantástico, puesto que lo maravilloso «está ahí, a cada paso [...] y consiste sobre todo en el hecho de que las pautas de la lógica [...] todo lo que nuestra inteligencia acepta desde Aristóteles como inamovible [...] por una especie de viento interior, que los desplaza y que los hace cambiar» (Cortázar, 1982). La nostalgia que se abre paso en el cuento se ajusta a la definición cortazariana del *sentimiento de lo fantástico* porque a pesar de no tener nada en común con «las grandes imaginaciones del cine», altera la realidad del protagonista.

Para dar cuenta de esta alteración conviene prestar atención a la técnica narrativa porque constituye el mecanismo utilizado por el autor primero para forjar el dualismo entre la realidad y el recuerdo y luego para difuminar los límites entre estas dos dimensiones. El relato se inicia y se finaliza con un narrador intradiegético autodiegético en primera persona: Aníbal es el escritor de su propio cuento. Entre ambas intervenciones de este narrador emerge una historia contada por un narrador extradiegético de omnisciencia selectiva, por el cual conocemos los sentimientos y los hechos ocurridos a Aníbal. «Se trataría, entonces, de un narrador que se desdobla en personaje que escribe y vive su propia ficción» (Sae-Yie, 2005) y el cuento se queda atrapado en el marco de la vida cotidiana del narrador. Sin embargo, no hay manera de contener la nostalgia en estos límites, porque el protagonista se da cuenta de la paradójica ausencia de vida en su día a día al que llama «presente utilitario sin recuerdos ni sueños» (Cortázar, 2003: 1067), afirmación que nos lleva a preguntarnos: ¿cuál es el tiempo verdadero del cuento?

Con el desdoblamiento del narrador en el autor y el protagonista de su propio cuento se desdobla también el tiempo y surge un segundo mundo: el mundo de los recuerdos, las

ilusiones y los sueños. De este modo, podemos postular la existencia dentro del cuento de dos niveles narrativos: el de la situación narrativa del protagonista y el de sus recuerdos. Es en el segundo nivel donde la nostalgia planta una semilla de lo fantástico que se expande hasta tal punto de que quepa la posibilidad de que el encuentro de Aníbal con Sara -en el que ella confiesa corresponderle en su amor- no pertenezca el terreno del recuerdo, sino sea una ensoñación. Sería simplista tachar de ilusiones y meras fantasías toda la historia contada en *Deshoras* y la llamada de Cortázar a «no aceptar otro orden que el de las afinidades, otra cronología que la del corazón, otro horario que el de los encuentros a deshora, los verdaderos» (Cortázar, 1984) hace que desechemos completamente esta concepción del cuento.

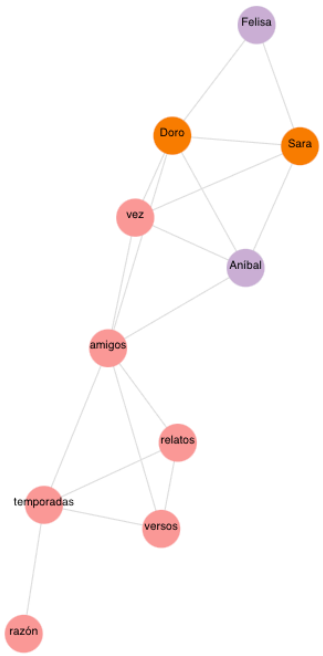

Figura 4. Subgrafo fuertemente conexo de *Deshoras*

El grafo de conocimiento extraído del texto de *Deshoras* es un grafo general que contiene un componente fuertemente conexo formado por un conjunto de 10 vértices (fig. 4). Este subgrafo engloba los nodos correspondientes a los personajes del cuento y establece vínculo entre ellos y la realidad inmediata del narrador, es decir, con el primer nivel narrativo del cuento. Este fenómeno ilustra la imposibilidad de que la nostalgia y el extrañamiento estén encerrados en el marco del ejercicio de escritura del narrador autodiegético, porque de los dos mundos presentes en *Deshoras*, solo el de la situación narrativa puede considerarse cerrado.

Para entender mejor este razonamiento, nos apoyaremos en la teoría del cuento que Cortázar expresa en su ensayo *Algunos aspectos del cuento* (1970). En este ensayo el escritor

describe los pasos que llevan a la creación de un cuento que se pueda calificar de *bueno*. Según él, un buen cuento es significativo, es decir, es un cuento que «quiebra sus propios límites con esa explosión de energía espiritual que ilumina bruscamente algo que va mucho más allá de la pequeña y a veces miserable anécdota que cuenta» (Cortázar, 1970). Todos los conceptos que definen el primer nivel narrativo de *Deshoras* están interconectados de tal forma que no queda ningún cabo suelto, ninguna posibilidad de que las ideas propias de la dimensión de «la pura vida sin más ni menos» (*Deshoras*) se proyecten en el lector y causen algún impacto significativo sobre él, a diferencia de las imágenes vívidas y variadas de la otra dimensión, mucho más amplia. Asimismo, el subgrafo fuertemente conexo ilustra la dualidad de las figuras de los personajes cuyos nodos vinculan los dos mundos: «en este sentido, *Deshoras* es una historia de dos personajes de dimensiones distintas que sin el poder de las palabras nunca habrían podido encontrarse de verdad» (Sae-Yie, 2005).

El peculiar tratamiento que hace Cortázar de tiempo en su cuento es otra herramienta de la que se sirve para construir y confundir los dos mundos en los que se desdobla el universo de *Deshoras*. La incertidumbre del tiempo fomenta la incertidumbre de la realidad y está presente a lo largo de todo el cuento. Si atendemos al primer nivel narrativo, el tiempo interno del texto dura aproximadamente una hora, hecho que se aclara en el último párrafo, cuando el narrador constata que «las palabras llevaban ya una hora [...] tendiéndose en el papel» (Cortázar 2003: 1067). En cambio, el tiempo en el que transcurre la historia subyacente al segundo nivel es mucho más amplio dado que abarca el período desde la infancia de Aníbal hasta su presente.

Dos elementos fundamentales para analizar la cronología de *Deshoras* son la noche y el sueño: los conceptos que se configuran como vínculos entre los dos niveles y contribuyen a la metamorfosis temporal. No conocemos desde el principio la hora a la que el protagonista se pone a escribir, solo en el último apartado se nos revela que es por la noche cuando le invade la nostalgia de su amor adolescente que sufría y disfrutaba de noche en la soledad de su cuarto de Bánfield. Además, este encuentro amoroso entre Sara y Aníbal del que se duda si es recuerdo o ensoñación se produce después de que él sueña con la chica.

El grafo conceptual evidencia con unas pruebas objetivas la relevancia de los conceptos del sueño y la noche. El nodo «noche» es el el vértice con el mayor valor de la centralidad de intermediación (2384.5), lo que señala su alta capacidad de influenciar sobre los demás vértices del grafo. Dicho valor es también muy alto para el nodo «sueño», equivale a 153.75 y

es casi tres veces mayor que la media entre todos los conceptos (57.12). La figura 5 muestra el fragmento del grafo que acoge todos los nodos que están directamente vinculados con noche y sueño. El tamaño del nodo refleja su valor de centralidad de intermediación y la oscuridad de la línea muestra el peso de la relación que se registró a partir de la frecuencia de su aparición a lo largo del texto. Conviene apuntar a que la relación entre Sara y noche se sitúa sexta según su peso, siendo este valor igual a nueve, lo que está por encima de la media (1.95). Muy significativa es la presencia en este fragmento del grafo de los elementos clave del segundo nivel narrativo: *patio*, como espacio predominante en los episodios de Bánfield, *cama* como el lugar en el que Aníbal enfermo sueña por primera vez con Sara, *pelo* como una de las pocas características físicas que se nos dan de Sara: «a veces Aníbal veía sobre todo su pelo rubio cayéndole sobre los hombros como una caricia que él hubiera querido sentir contra su cara» (Cortázar, 2003: 1056), etc. También podemos reparar en las palabras sentimentales y sensuales, como *delicia*, *caricia*, *dolor*, *fatiga* que hacen referencia al amplio espectro de emociones que puede producir el amor en una persona.

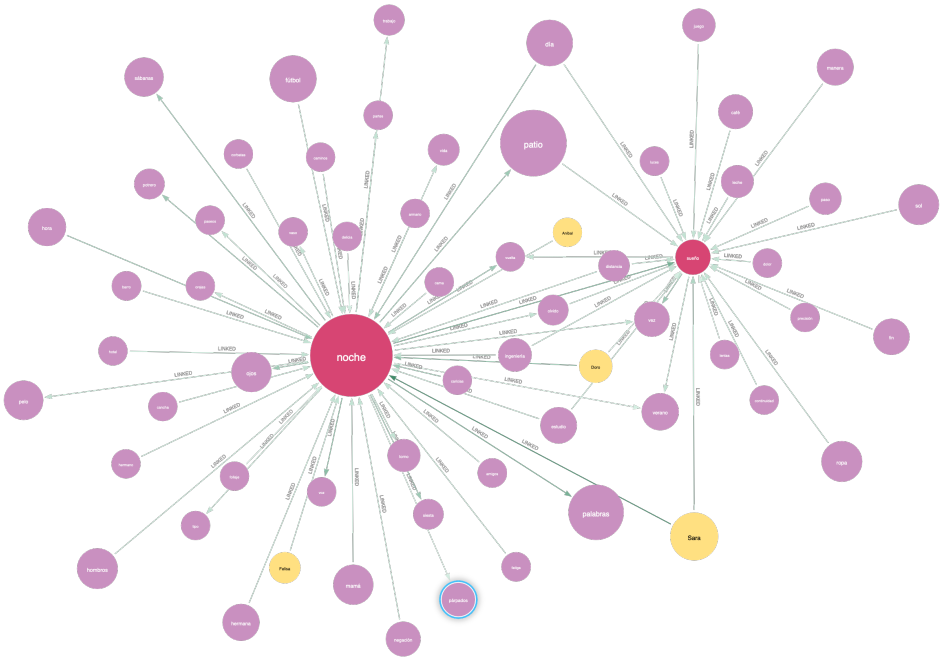

Figura 5. Los nodos que están en relación directa con *noche* y *sueño*

Todos estos datos llevan a la conclusión de que la oscuridad de la noche es una especie de punto inicial de todo el cuento, es el factor que origina la nostalgia, el elemento que potencia la confusión espacio-temporal, acentuada con expresiones de tipo «llegar a un tiempo», «llegar al hoy nuestro de cada día». Sin embargo, si afirmáramos que el desdoblamiento de la realidad del cuento se fundamenta en la oscuridad de la noche y la nostalgia, omitiríamos un

elemento fundamental que sirve de vínculo entre los dos mundos. Este elemento está presente en la Figura 5 y su relación con la noche está entre las diez más significativas de todo el grafo de conocimiento. Las palabras viajan por el tiempo y por el espacio uniendo los mundos y los destinos que no podrían encontrarse de otra forma. Las palabras hacen que el mundo de los recuerdos y las ilusiones parezca más real que el mundo real:

> ¿Para qué seguir escribiendo si las palabras llevaban ya una hora resbalando sobre esa negación, tendiéndose en el papel como lo que eran, meros dibujos privados de todo sostén? Hasta algún momento habían corrido cabalgando la realidad, llenándose de sol y verano, palabras patio de Bánfield, palabras Doro y juegos y zanjón, colmena rumorosa de una memoria fiel. Sólo que al llegar a un tiempo que ya no era Sara ni Bánfield el recuento se había vuelto cotidiano (Cortázar, 2003: 1067).

No obstante, la finalización de la escritura y el hecho de haber sumergido la mesa de trabajo en la oscuridad no cierran la puerta a los sentimientos, no rompe el vínculo establecido con el otro mundo, que se rige por la cronología de los encuentros a deshora. El último párrafo del cuento se caracteriza por oraciones muy largas con las que se acelera el ritmo de la narración y se intenta transmitir la desesperación que siente protagonista inconforme con que el destino lo había apartado de Sara. En este emocionante final la noche de encuentro con Sara se configura como un cronotopo que quintaesencia el otro mundo en el que la nostalgia se llena de vida, en el que los dos destinos que pasaron una al lado de otro sin encontrarse llegan por fin a juntarse. Aunque el protagonista tache de imposible esta otra vida, las palabras pueden hacerla parecer más real que la realidad misma. Esta vida va a continuar irrumpiendo en la realidad en forma de fantasías y sueños porque el sentimiento de lo fantástico es una manera de ver el mundo y en este sentido, podemos decir que Cortázar se proyecta a si mismo y a su manera de ver el mundo en su protagonista quien, al igual que el autor, siente «que entre dos cosas que parecen perfectamente delimitadas y separadas, hay intersticios por los cuales [...] pasaba, se colaba, un elemento, que no podía explicarse con leyes, que no podía explicarse con lógica, que no podía explicarse con la inteligencia razonante» (Cortázar, 1982).

El hecho de que Cortázar retrate el proceso creativo y llame la atención sobre el poder de la palabra puede verse como una vuelta al origen del texto en la escritura. Pero en *Deshoras* están cifrados también el origen del adulto en su infancia y el amor como el sentimiento primigenio. Según señala Sae-Yie (2005), *Deshoras* es «un cuento de amor que se hace posible gracias a la escritura de los recuerdos» y en nuestro análisis vimos cómo se hace posible la articulación de dos niveles narrativos que, interactuando entre sí e irrumpiendo en la realidad del lector, rinden homenaje al amor y al poder de las palabras.

4. CONCLUSIONES

No debemos incurrir en el error de que la interdisciplinariedad supone la pérdida de identidad de las disciplinas, homogeneizando el panorama científico que si sigue el camino en el que está ahora llegará a un saber absoluto y totalizador. Por el contrario, los proyectos interdisciplinares aprovechan la teoría y los métodos propios de distintas disciplinas con el fin de, sin perder su identidad, poder abarcar la totalidad del objeto de estudio.

La cultura es un asunto de perspectiva, un modo de ver cuya interpretación depende de la ideología, es decir, el punto de vista es el que le otorga sentido y construye una verdad. Con el desarrollo del presente trabajo pretendimos conocer las implicaciones sociales que tiene la Revolución Digital. El hecho de vivir en un mundo interconectado y desenvolverse en entornos tecnológicos ha digitalizado el código cultural de la humanidad.

Las Humanidades, como el conjunto de disciplinas que se encargan de estudiar al ser humano y su actividad social, se tienen que enfrentar actualmente a un importante desafío para poder abarcar la cultura multimodal. En este trabajo intentamos indagar en los principios fundamentales de las nuevas Humanidades, conocidas bajo la denominación Humanidades Digitales. Una revisión histórica de las trayectoria de la consolidación de esta disciplina permitió investigar sobre las necesidades y los retos originarios que demandaron la aplicación de metodología computacional a los datos humanísticos. La revolución iniciada con el *Index Thomisticus* de Roberto Busa desembocó en la reinterpretación de las Humanidades como práctica generativa y no solamente epistémica.

La casi infinita libertad creativa que ofrecen los entornos digitales para la creación y la investigación hace que no sea posible ni necesario establecer fronteras rígidas tanto de las Humanidades Digitales como de su dimensión filológica. Sin embargo, uno de los objetivos del marco teórico de la presente investigación fue trazar el recorrido compuesto de tales pasos esenciales como curación, análisis, edición y modelado, que se tienen que seguir para adaptar

la metodología humanística a los entornos computacionales, cuya principal limitación consiste en la intolerancia hacia la ambigüedad en los niveles primarios de cálculo. El proceso de la elaboración de una edición digital nos sirvió para ilustrar el hecho de que la flexibilidad que ofrece la nueva metodología obliga al humanista digital a asumir responsabilidad sobre la adecuada elección de formato maestro del proyecto, el tipo de marcado textual y la configuración de la licencia bajo la cual se distribuirán los datos. De este modo, la ausencia de una manera estrictamente correcta de proceder se convierte en una de las características de la nueva disciplina. Así se cumple la predicción de Roberto Busa: los humanistas digitales tienen la oportunidad de mostrar su creatividad en casi ausencia de límites.

Por otro lado, uno de los propósitos de la investigación fue comprobar que el entorno digital, como cualquier espacio de interacción, impone cierto tipo de límites que derivan de la necesidad de compatibilizar los diferentes recursos con las herramientas y programas existentes. Para facilitar creación de recursos reutilizables se han desarrollado estándares, como el TEI, que, sin limitar demasiado la libertad de los creadores, hacen posible que los datos generados como resultado de una investigación encajen bien en el ecosistema tecnológico.

Además de nuevas técnicas de estructuración, almacenamiento y manejo de datos, las Humanidades Digitales abren nuevas vías para el análisis literario. El texto es un hecho cultural, y consecuentemente social, en el que está cifrado un esfuerzo de reflexión sobre la realidad que se desenvuelve ante los sentidos de los lectores. Las palabras unen los respectivos sistemas conceptuales del autor y del lector transmitiendo significados, ideas y sentimientos que pasan por el filtro de los sesgos personales del destinatario para ser interpretados. Este hecho puede dar lugar a infinitas interpretaciones de los textos que se construyen, sin embargo, en ausencia de un fundamento objetivo.

Tener una oportunidad de conocer una abstracción del sistema lingüístico empleado por el autor en ausencia de todos los factores que condicionan la interpretación de el texto a través del filtro sentimental significa abrir una nueva perspectiva hacia la obra estudiada. En este trabajo presentamos un método para la construcción de un grafo del conocimiento mediante la aplicación de tecnologías de PLN al texto del cuento *Deshoras* de Julio Cortázar. Con la aplicación de la teoría de los grafos la abstracción del sistema conceptual del cuento pretendimos proporcionar una base objetiva para el análisis literario, esclareciendo las

relaciones existentes entre los personajes, los conceptos y los dos mundos que construye Cortázar.

Conviene notar que la metodología presentada puede utilizarse para investigaciones con planteamiento distinto al nuestro. En este trabajo el método cuantitativo se conjuga con el análisis tradicional, sirviendo de pruebas para las hipótesis informativas. Pero todos los datos cuantitativos extraídos y por extraer del grafo obtenido pueden ser analizados en escenarios de lectura distante cuando el objetivo es indagar en las características de un conjunto de cuentos. Por lo tanto, podemos afirmar que el trabajo presentado aspira cumplir con el criterio de sostenibilidad que imponen las premisas teóricas de las Humanidades Digitales en tanto que el método puede ser aplicado a otros textos o corpus.

5. BIBLIOGRAFÍA

Boldi, P., Vigna, S. (2014): "Axioms for Centrality", *Internet Mathematics*, XIX, pp. 222-262.

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T., & Schnapp, J. (2012): *Digital_Humanities*. Mit Press.

Busa, R. (1980): "The annals of Humanities Computing: the Index Thomisticus", *Computers and the Humanities*, 14, pp. 83-90.

Calvo Tello, J. , Gómez, S., González, J. M., & Vilches, R. (2015): "Hacia una biblioteca electrónica textual del teatro en español de 1868-1936 (BETTE)", *Texto Digital*, XI, 2, pp. 171-184.

Calvo Tello, J. (2019): "Diseño de corpus literario para análisis cuantitativos", *Revista de Humanidades Digitales*, 4, pp. 115-135 <enlace>.

Chizhikova, M. *et al* (2021): "SINAI at BioCreative VII LitCovid Track: Corpus augmentation for COVID-19 literature multi-label classification", *Proceedings of the BioCreative VII Challenge Evaluation Workshop*, VV.AA., pp. 336-340 <enlace>.

Corral, M. D. (2009): "La Biblioteca Digital Hispánica", en *IX Workshop Rebiun de proyectos digitales. Investigación e información: tendencias en los sistemas digitales de gestión de la producción científica*, VV.AA., Salamanca, Universidad de Salamanca.

Cortázar, J (1984): "Para leer de forma interrogativa", en *Salvo el crepúsculo*, Madrid, Alfaguara. pp. 39-41.

Cortázar, J. (1969): "Del cuento breve y sus alrededores", *Último round*, 1, 59-82.

Cortázar, J. (1982): "El sentimiento de lo fantástico", en *La vuelta al día en ochenta mundos*, Buenos Aires, Siglo veintiuno editores, pp. 43-44.

Cortázar, J. (1983): "La esfera de los cuentos / entrevistado por José Julio Perlado", *Espéculo*, 2, s.p. <enlace>.

Cortázar, J. (2003): “Deshoras”, en *Obras completas I. Cuentos*, ed. de S. Yurkievich, Barcelona, Galaxia Gutenberg, pp. 1054-1068.

Ehrlicher, H., Lehmann (2020): “A Quantitative View of Genre Poetics. Calderón de la Barca’s oeuvre and the contemporaneous poetics of drama”, *LitLab Pamphlets*, 9 <enlace>.

Euler, L. (1741): “Solutio problematis ad geometriam situs pertinentis”, *Commentarii academiae scientiarum Petropolitanae*, pp. 128-140 <enlace>.

Fischer F. *et al* (2017): “Network Dynamics, Plot Analysis. Approaching the Progressive Structuration of Literary Texts”, *DH2017: 8–11 August 2017. Book of Abstracts*, McGill University, Montréal <enlace>.

Fischer, F. *et al* (2019): “Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama”, *Proceedings of DH2019: "Complexities"*, Utrecht University <enlace>.

Fradejas Rueda, J. M. (2009-2010): “La codificación XML/TEI de textos medievales”, *Memoriabilia*, 2, pp. 219-247.

Freeman, L. (1977): “A set of Measures of Centrality Based on Betweenness”, *Sociometry*, 40/1, pp. 35-41.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). *Learning word vectors for 157 languages*. arXiv preprint arXiv:1802.06893. <enlace**>.**

Haslhofer, B., Isaac, A., Simon, R. (2018): *Knowledge graphs in the libraries and digital humanities domain*. [arXiv preprint arXiv:1803.03198] <enlace>.

Maximova, D., Fischer, F. (2018): “A Quantitative Study of Stage Directions in Russian Drama”, *EADH2018: «Data in Digital Humanities»*, Galway, National University of Ireland.

Mendenhall, T. C. (1901): “A mechanical Solution of a Literary Problem”, *Popular Science Monthly*, LX, pp. 97-105 <enlace>.

Menéndez Velázquez, A. (1998): “Una breve introducción a la teoría de grafos”, *Suma*, 28, pp. 11-26 <enlace>.

Meneses, S., Elena, E. (2021): “Fundamentos neurocognitivos como sustrato para un diseño intuitivo”, *Revista Humanidades*, XI, 1, pp. 39-55 <enlace>.

Michel J. B., Lieberman Aiden A., *et al*., (2010): “Quantitative Analysis of Culture Using Millions of Digitized Books”, *Science*, CCCXXXI, 6014, pp. 176-182 <enlace>.

Moreno, J. L. (1941): “Foundations of sociometry: An introduction”, *Sociometry*, IV, 1, pp. 15-35.

Mosteller, F., Wallace, D. (1963): «Inference in an Authorship problem», *Journal of the American Statistical Assosiation*, LIV, 302, pp. 275-309 <enlace>.

Qin, Z. et al. (2020): “Multitask mixture of sequential experts for user activity streams”, *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ACM, New York pp. 3083-3091 <enlace>.

Rojas Castro, A. (2013): “Las Humanidades Digitales: principios, valores y prácticas”, *Janus. Estudios sobre el siglo de oro*, 2, pp. 74-99 <enlace>.

Sae-Yie, P. (2005): “Un encuentro a deshoras: Discurso y tiempo en *Deshoras*, de Julio Cortázar”, *Espéculo. Revista de estudios literarios*, 29 <enlace>.

Salton, G., & Buckley, C. (1988): “Term-weighting approaches in automatic text retrieval” En *Information processing & management*, XXIV, 5, pp. 513-523 <enlace>.

Singhal A. (2012): “Introducing the knowledge graph: Things, not strings”, *Google. The keyword* <enlace>.

Smith, D., Rydberg-Cox, J., Crane G., «The Perseus Project: a digital library for the humanities», *Literary and Linguistic Computing*, XV, 1, pp. 15–25.

Taulé, M., Martí M.A., M. Recasens (2008): «Ancora: Multilevel Annotated Corpora for Catalan and Spanish», en *Proceedings of 6th International Conference on Language Resources and Evaluation*, Calzolari, N *et al* (Eds.), Marrakesh, Morocco, pp. 797-802.

TEI Consortium (Eds.) *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* v. 4.4.0. <enlace>.

Wendell, I. (2021): *A Statistical Analysis of Genre Dynamics: Evolution of the Russian Five-Act Comedy in Verse in the Eighteenth and Nineteenth Centuries* [Tesis doctoral], University of California, Los Angeles.