PUBLICACIONES DE LA
ASOCIACIÓN ARGENTINA
DE HUMANIDADES DIGITALES

Análisis distante en un corpus bilingüe: el caso de Rosalía de Castro

Distant Analysis in a Bilingual Corpus: The Case of Rosalía de Castro

Rocío Luciana MÉNDEZ

mendezzrocio@gmail.com

Universidad de Buenos Aires

RESUMEN

En este trabajo nos dedicaremos a compartir nuestra experiencia con la herramienta Voyant Tools sobre la obra poética, tanto en español como en gallego, de Rosalía de Castro (1837-1885). Esta reseña fue realizada para la Diplomatura en Humanidades Digitales (UCES) dirigida por Gimena del Rio Riande (Consejo Nacional de Investigaciones Científicas y Técnicas). Nos interesa desarrollar un estudio de caso y experiencia de usuario aplicando tecnologías del lenguaje y explicando las ventajas y desventajas de trabajar con un corpus bilingüe, por lo que el siguiente trabajo trata de una descripción de la herramienta en uso en donde se analizarán distintas variables observadas en la aplicación de las diferentes funciones disponibles en Voyant Tools. En términos generales, buscamos generar un primer acercamiento que favorezca una aproximación al análisis distante de un corpus bilingüe.

PALABRAS CLAVE

Rosalía de Castro, Voyant Tools, lectura distante, corpus bilingüe, tecnologías del lenguaje.

ABSTRACT

Our proposal is to share our experience with Voyant Tools on the poetic work, both in Spanish and Galician, by Rosalía de Castro (1837-1885). This review was made for the Diplomatura en Humanidades Digitales (UCES) directed by Gimena del Rio Riande (National Council for Scientific and Technical Research). We are interested in developing a case study and user experience when working with language technologies, explaining the opportunities and difficulties that we perceived when working on a bilingual corpus, therefore the following work is a description of the tool in use where different variables will be analyzed with the different functions available in Voyant Tools. In conclusion, we seek to generate a first approach that benefits an approximation in distant analysis in a bilingual corpus.

KEYWORDS

Rosalía de Castro, Voyant Tools, distant reading, bilingual corpus, language technologies.

PublicAAHD - 2020

Vol. 1

ISSN

2718-7470

Publicaciones de la Asociación Argentina de Humanidades Digitales PublicAAHD - 2020 Vol. 1

https://doi.org/10.24215/27187470e007

1. INTRODUCCIÓN

Para el siguiente trabajo se utilizó la herramienta Voyant Tools[1], desarrollada por Stéfan Sinclair (McGill University) y Geoffrey Rockwell (University of Alberta). Esta aplicación de código abierto, implementada en trabajos de investigación y de divulgación basados en el análisis de grupos textuales, ayuda a generar nueva información y conocimiento oculta en la *big data*. En las siguientes secciones, se aplicarán algunas de las posibilidades analíticas[2] de Voyant Tools para la lectura y el análisis textual, tomando como punto de partida el conjunto de poemarios escritos por Rosalía de Castro (1837-1885). Para quienes comienzan desde cero el uso de esta herramienta, se recomienda la consulta de la guía de uso en la web de la aplicacion y el tutorial en español realizado por Silvia Gutiérrez para el proyecto de Humanidades Digitales The Programming Historian[3].

El corpus elegido comprende dos libros en gallego, *Cantares gallegos* (1863), *Follas novas* (1880), y tres libros en castellano, *La Flor* (1857), *A mi madre* (1863) y *En las orillas del Sar* (1884). La escritora compostelana es una de las figuras más importantes del *Rexurdimento*[4] gallego, cuya misión fue recuperar el prestigio que supo tener la lengua gallega mediante una escritura que aborda la temática histórica, folklorista y costumbrista característica de la lírica gallegoportuguesa. Si tomamos el concepto de conciencia lingüística[5], comprenderemos que la literatura consolida el reconocimiento del idioma por el respeto cultural que esta provoca, de esta forma resulta de especial interés el estudio del conjunto seleccionado, siendo el propósito de esta reseña un aspecto fundamental de la obra poética de Rosalía de Castro: el bilingüismo.

El objetivo es plasmar una serie de observaciones realizadas en torno a la poesía de Rosalía de Castro desde la perspectiva del estudio computacional de una obra literaria bilingüe y, en especial, de la visualización de la información, con el fin de compartir algunas de las posibilidades que ofrece el hecho de poder trabajar con grandes cantidades de datos derivados de las obras literarias.

2. ANÁLISIS DE CORPUS

Los datos cuantitativos y herramientas de análisis de contenido aportan información sobre la existencia de patrones léxicos y discursivos en las obras literarias. La función principal de Voyant Tools está basada en el estudio de las palabras más frecuentes que pueden contribuir al reconocimiento de estructuras recurrentes en la obra de los autores.

---

[1] Accesible desde: https://voyant-tools.org/.
[2] Las funcionalidades de *Voyant Tools* son muy variadas y pueden consultarse aquí: http://docs.voyant-tools.org/tools/.
[3] Accesible desde: https://programminghistorian.org/es/lecciones/analisis-voyant-tools.
[4] *Rexurdimento* es el nombre con el que se conoce al período cultural sucedido en Galicia en el siglo XIX y que tuvo como principal ambición la revitalización de la lengua y la identidad gallega.
[5] El término conciencia lingüística consiste en "el conocimiento explícito acerca de la lengua y la percepción y sensibilidad conscientes al aprender la lengua, al enseñarla y al usarla" *(Association for Language Awareness)*, es decir, en lo referido al aprendizaje, permite advertir diferentes aspectos de la lengua que de otro modo pasarían inadvertidos.

Una vez cargados los poemarios, Voyant Tools nos ofrece un resumen general de los datos[6] que ha extraído del grupo. El primer dato destacable es la mayor extensión de las obras en gallego frente al de las obras en castellano (figura 1).

Figura 1. Extensión del corpus. Elaboración propia.

Al utilizar la visualización *Documentos*, notaremos que *Follas novas* es la obra más larga de la autora, pues es la que tiene más palabras con un total de 23.396, siendo 6.057 palabras tipo o palabras únicas (figura 2).

Sumario | Documentos | Frases

	Título	Palabras	Tipos	Proporción	Palabras/Oración
1	1857 La Flor	5,226	1,621	31%	20.2
2	1863 A mi madre	2,015	779	39%	15.3
3	1863 Cantares gallegos	18,187	4,738	26%	19.2
4	1880 Follas novas	23,396	6,057	26%	18.0
5	1884 En las orillas del Sar	16,589	3,831	23%	24.9

Figura 2. Palabras únicas y palabras tipo. Elaboración propia.

Sin embargo, en proporción, *A mi madre* es la obra que posee el lenguaje más variado de todo el corpus, a pesar de ser el poemario más corto. Con tan sólo 2.015 palabras y 779 palabras únicas, posee el mayor porcentaje de densidad de vocabulario[7] del grupo (figura 3).

Figura 3. Densidad del vocabulario del corpus. Elaboración propia.

Si observamos los datos presentados podríamos inferir que los poemarios más largos, al demostrar poca densidad de vocabulario, se vuelven más reiterativos y las palabras comienzan a ser redundantes entre sí, dando a suponer que son acotadas las temáticas que se repiten a lo largo de dichos poemas.

Una vez excluidas las *stopwords*, las palabras más frecuentes que integran el corpus son *dios* (166), *eu* (145), *vida* (138), *triste* (136), *sol* (124). Al trabajar con un grupo de textos bilingües, la herramienta *Tendencias* nos permite notar que hay términos que están presentes en ambas lenguas, como lo son *dios, vida, triste* y *sol*, ya que se encuentran en los cinco elementos del conjunto (figura 4).

[6] Para experimentar con el corpus poético de Rosalía de Castro: https://voyant-tools.org/?corpus=ad510111ddb802ca9d6228629a990d23.

[7] La densidad de vocabulario se obtiene al dividir el número de palabras únicas entre el número de palabras totales. Cuanto más alto es el índice de densidad significa una mayor variedad de palabras.

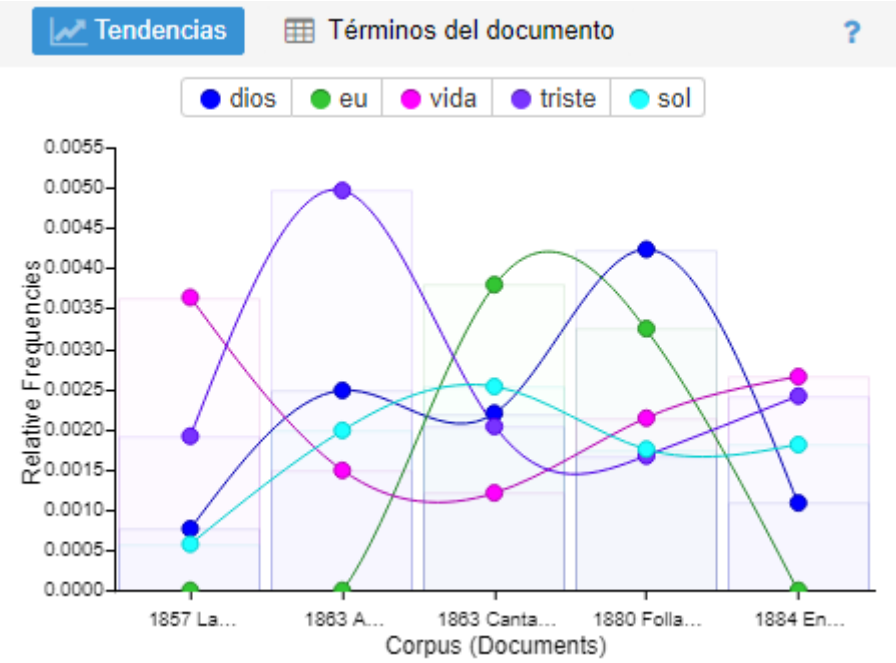

Figura 4. Tendencias de las palabras más frecuentes del corpus. Elaboración propia.

La gráfica revela que el *sol* es un elemento importante a lo largo de la obra de Rosalía, mientras que el término *triste* muestra una tendencia mayor en el poemario *A mi madre*, obra que escribe luego del fallecimiento de su madre.

Con respecto a las 145 apariciones del pronombre personal *eu* (yo), podemos notar una fuerte presencia del *yo* en el corpus gallego frente a las 86 apariciones del pronombre en castellano. Al detenernos en el contexto de la palabra notamos frases que de cierta manera reafirman la postura de Rosalía frente al uso de la lengua de su tierra, Galicia: "[...] na lengua qu' eu falo" (De Castro, 1863, I-30), pone en su voz el sentimiento de tristeza: "[...] eu che cantaba en triste soledá" (XII, I, 33) y reflexiona sobre su rol de escritora: "[...] as cousas qu'ora eu penso. E ben, ¿para qu'escribo?" (1880, I, II, 4). De las cuales podemos obtener una lectura más amplia del entorno en la ventana Lector en el que se halla la palabra y poder comprender mejor su sentido (figura 5).

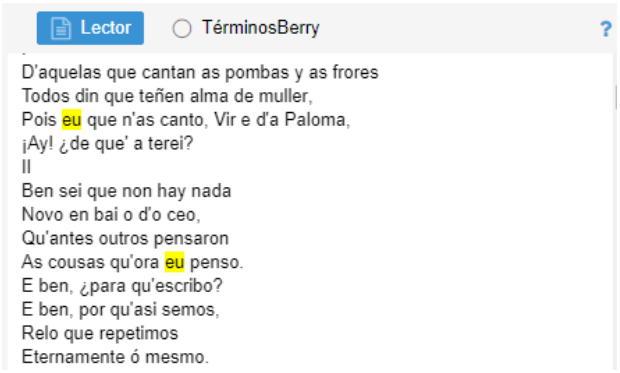

Figura 5. Vista de Lector del poema Follas novas. Elaboración propia.

Las *Líneas de burbuja* visualizan la frecuencia y repetición del uso de una palabra en un corpus. Cada palabra seleccionada se representa como una burbuja donde el tamaño de esta indica la frecuencia de la palabra en el segmento de texto correspondiente (figura 6).

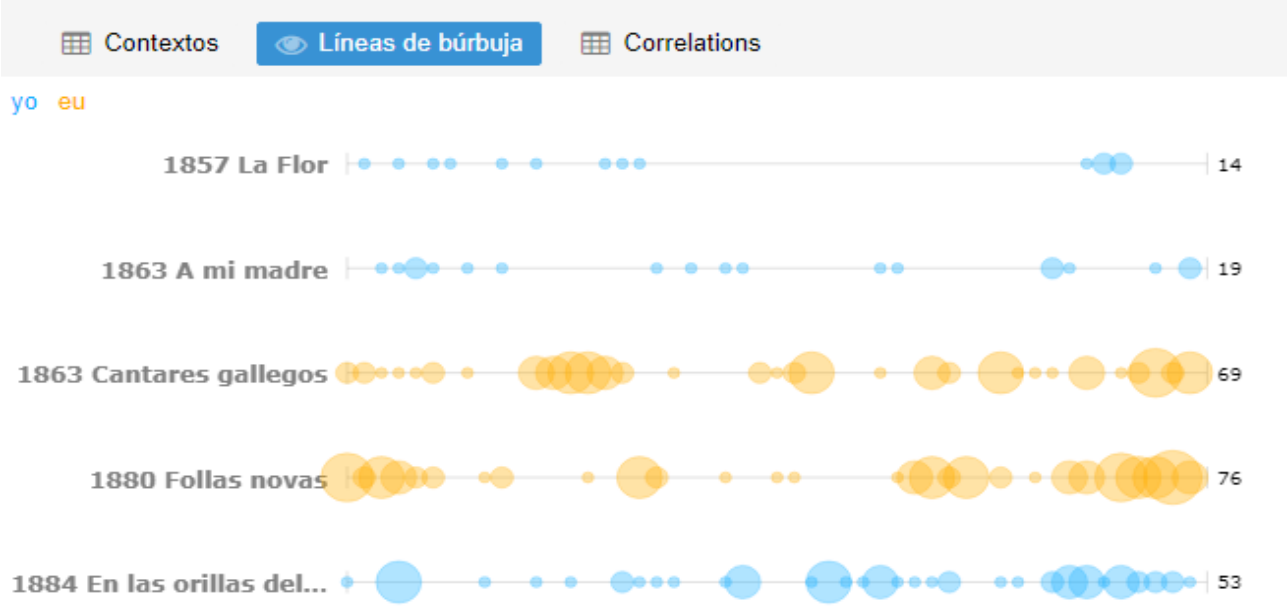

Figura 6. Líneas de burbuja con los términos eu/yo. Elaboración propia.

En cuanto a las palabras diferenciadas de cada poemario en comparación con las restantes obras del grupo, la herramienta Voyant Tools nos arroja el siguiente resultado:

> Palabras diferenciadas (comparado con el resto del corpus):
> La Flor: flor (27), inés (12), joven (17), delicia (5), día (15).
> A mi madre: yo (19), madre (18), tuve (4), nieblas (3), hilo (3).
> Cantares gallegos: eu (69), noite (39), airiños (21), vexo (19), meniña (31).
> Follas novas: eu (76), adios (32), terra (55), dia (28), tí (25).
> En las orillas del Sar: yo (53), viejo (11), fuente (19), tierra (32), blanca (16).

A partir de estos datos, al analizar un grupo textual bilingüe, damos cuenta de los términos más significativos de cada poemario, especialmente de cada lengua y quizás inferir la temática y tintes emocionales con las que se expresa.

En *Cantares gallegos* apreciamos una particularidad del lenguaje gallego como lo es el diminutivo (*airiños*). En la siguiente figura, vemos cómo podemos encontrar los diminutivos en el texto con la *búsqueda sintáctica* de los términos coincidentes con los diminutivos *-iño, -iña, -iños, -iñas*:

#	Términos	Contar	Relativo	Tendencia
3	.*iña	273	15,011	
3	.*iño	169	9,292	
3	.*iñas	124	6,818	
4	.*iña	137	5,856	
3	.*iños	97	5,333	
4	.*iño	87	3,719	
4	.*iñas	31	1,325	
4	.*iños	29	1,240	

Figura 7. Búsqueda sintáctica por coincidencia de términos. Elaboración propia.

En una búsqueda por contexto podremos notar que son comunes los diminutivos de sustanti-

vos, como, por ejemplo, *auguiña, paxariños, casiña, airiños, olliños, mañanciña, anxeliños, campaniñas*. Esta observación, frente a la forma diminutiva más común perteneciente al castellano: *-ito, -ita, -itos, -itas* denota la nula existencia de los diminutivos en los poemas castellanos. La profusión de estas formas en el corpus gallego demuestra una clara decisión de Rosalía de reproducir las formas populares del habla de los habitantes de Galicia.

3. ANÁLISIS INDIVIDUAL DENTRO DEL CONJUNTO DE OBRAS SELECCIONADAS

En Voyant Tools también podemos dirigir nuestro enfoque hacia un estudio individualizado de las diferentes obras sin necesidad de cargar nuevamente el texto a la página web. En este caso tomaremos el poemario *La Flor* para llevar a cabo un breve comentario. Si nos detenemos en las palabras diferenciadas, notamos que el término *Inés* es el único antropónimo que destaca en el corpus poético general. El siguiente término que aparece con recurrencia es *joven*, en una primera sospecha podríamos imaginar que se trata de un adjetivo usado para describir al personaje principal, sin embargo, al observar el contexto del término advertimos que se refiere a un personaje masculino y únicamente en cuatro oportunidades se refiere a un personaje femenino.

Contextos | Líneas de búrbuja | Correlations

Documento	Izquierda	Términos	Derecha
1) 1857 L...	la sombra fresquísima escondidas. ...	joven	allí inmóvil descansaba cabe del
1) 1857 L...	tarde, amortiguado y yerto, aquel	joven	tal vez lo recogía... Clavado
1) 1857 L...	mecía, y en vano el	joven	revivir la quiere. Y también
1) 1857 L...	su reposo, paróse junto al	joven	que extasiado mirándole en su
1) 1857 L...	entre la mano... ...	joven	palabras pronuncia, que él sólo
1) 1857 L...	que se dormía, con el	joven	aquél, en los vapores que

Figura 8. Contexto del término joven en La Flor. Elaboración propia.

Al dirigir nuestro análisis hacia los términos más llamativos que aparecen en la nube de palabras que arma Voyant Tools:

Figura 9. Nube de palabras del poema La Flor. Elaboración propia.

Seleccionamos las palabras más significativas para obtener una visualización con la herramienta *Tendencias*. Del gráfico arrojado, uno podría intuir que la autora divide su obra en tres partes, la primera evocando al *amor* junto con el *dolor*, luego presenta al joven que seducirá a nuestra protagonista.

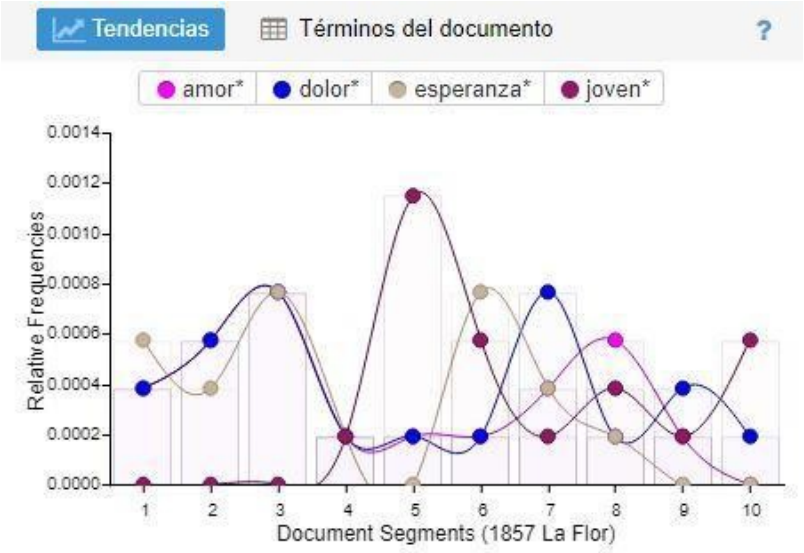

Figura 10. Tendencias en La Flor. Elaboración propia.

Cuando la autora introduce al personaje del joven en el poemario cerca de la mitad de la obra, los términos *esperanza, dolor* y *amor* decrecen. Luego, cuando el joven se encuentra ausente en el poema, aumenta la tendencia del término *dolor*, lo que podría denotar el sentimiento de la protagonista frente a la ausencia del amado (segmento 7).

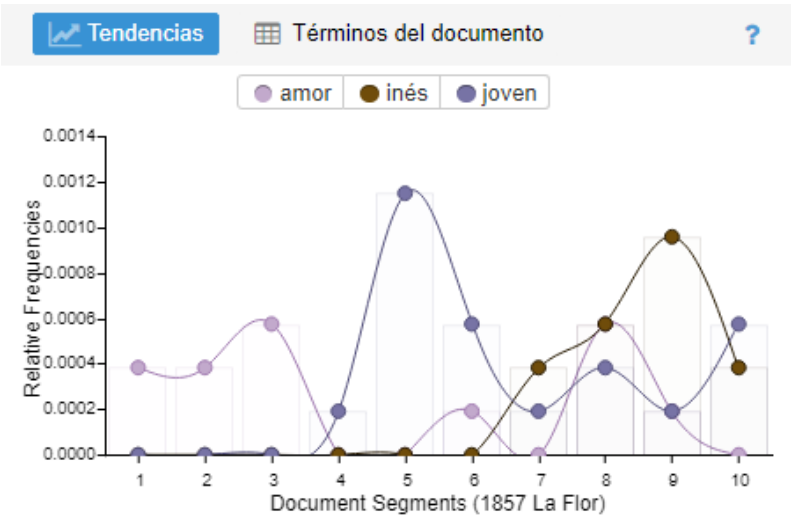

Figura 11. Tendencias de términos escogidos en el poema La Flor. Elaboración propia.

Estas gráficas pueden darnos una idea de los temas que predominan a lo largo de la obra. La tradición lírica puede suponer la predominancia de la aparición de la figura masculina hacia el centro del poemario y durante dos terceras partes del texto. Dicha irrupción es importante al tratarse de un poema que trata del amor, tanto es así que ante la ausencia del amado incrementa la presencia de los términos *amor, dolor* y *esperanza* (figura 10) y deja relegada la figura de Inés hacia el último tercio de la obra donde el término *joven* y *amante* caen en la Tendencia y aumenta el sentimiento relacionado a la tristeza (figura 12).

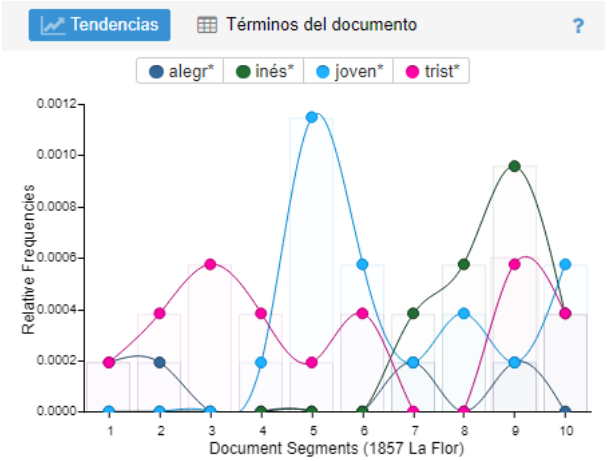

Figura 12. Tendencias de términos escogidos en el poema La Flor. Elaboración propia.

4. COMPARACIÓN ENTRE OBRAS

Una de las herramientas comparativas que ofrece Voyant Tools es *Scatterplot*, que nos permite comparar los estilos de un escritor o escritores a través de visualización gráfica de cómo los documentos de un conjunto se relacionan entre sí mediante la función similitud de documentos, análisis de correspondencias o análisis de componentes principales.

En las siguientes pruebas hemos empleado diferentes modos de examen de datos, los cuales permiten estudios muy provechosos debido a las diferentes posibilidades combinatorias. Si tomamos la frecuencia absoluta, claramente, al tratarse de un corpus bilingüe notamos a simple vista esta lejanía en las obras. No obstante, los resultados también muestran que ambos poemarios en lengua gallega tienen mayor cercanía en comparación a las obras en castellano (figura 13).

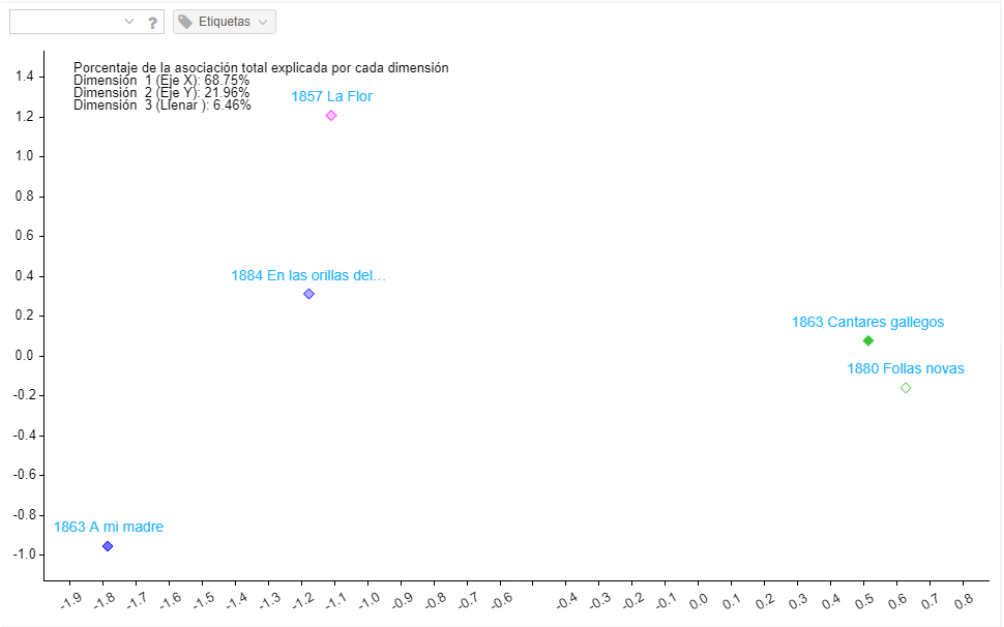

Figura 13. Similitud del corpus midiendo la frecuencia inversa de términos. Elaboración propia.

Si tomamos en cuenta ahora la frecuencia absoluta, estableciendo un parámetro de búsqueda de los 55 términos más habituales del conjunto, y observamos la similitud existente entre los documentos, el resultado es una distancia mayor entre *La Flor y A mi madre* que se encuentra más cercana a *En las orillas del Sar* (figura 14).

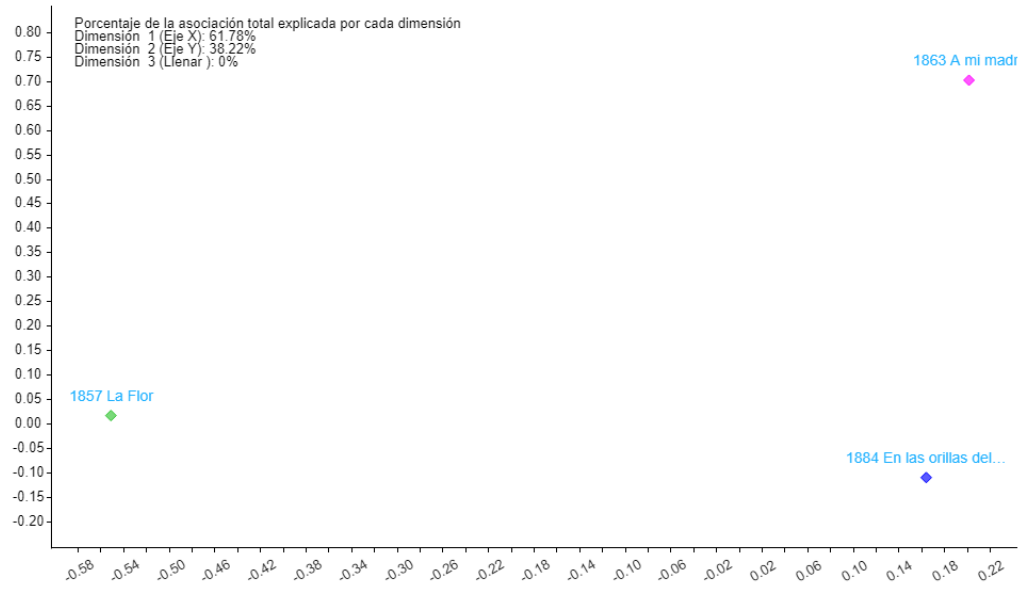

Figura 14. Similitud del corpus castellano midiendo las frecuencias absolutas. Elaboración propia.

Aplicando el mismo procedimiento, observamos que *Cantares gallegos* y *Follas novas* aún permanecen cercanas (figura 15).

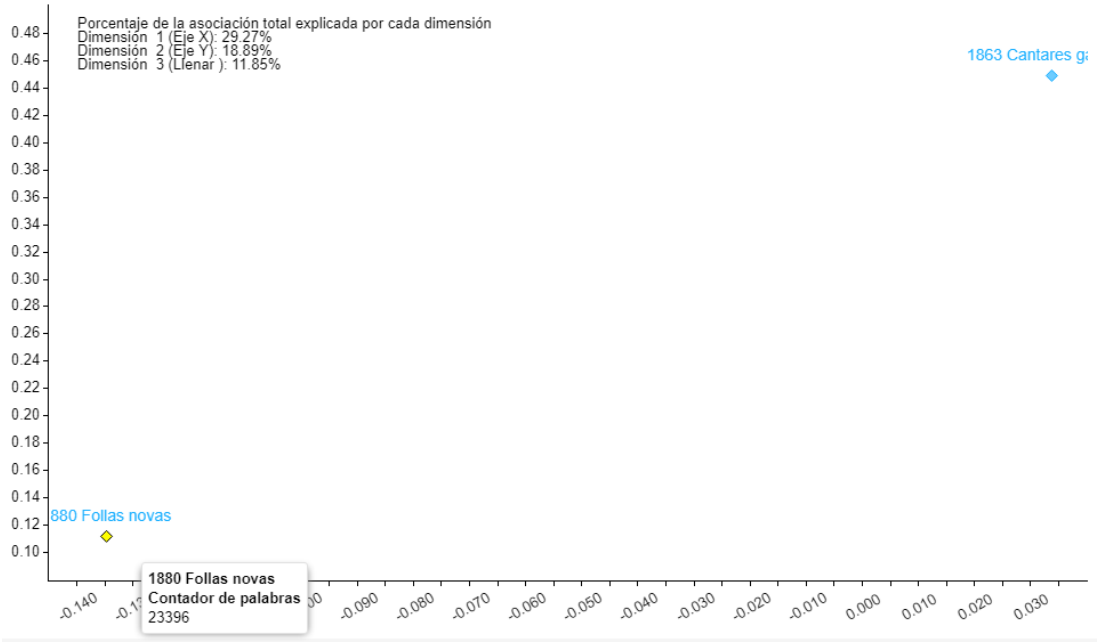

Figura 15. Similitud del corpus gallego midiendo las frecuencias absolutas. Elaboración propia.

Aplicando el análisis de correspondencias midiendo la frecuencia absoluta de los términos más usados en el corpus, observamos que, por ejemplo, la palabra *dios* se encuentra más cercana a la obra *Follas novas* (figura 16).

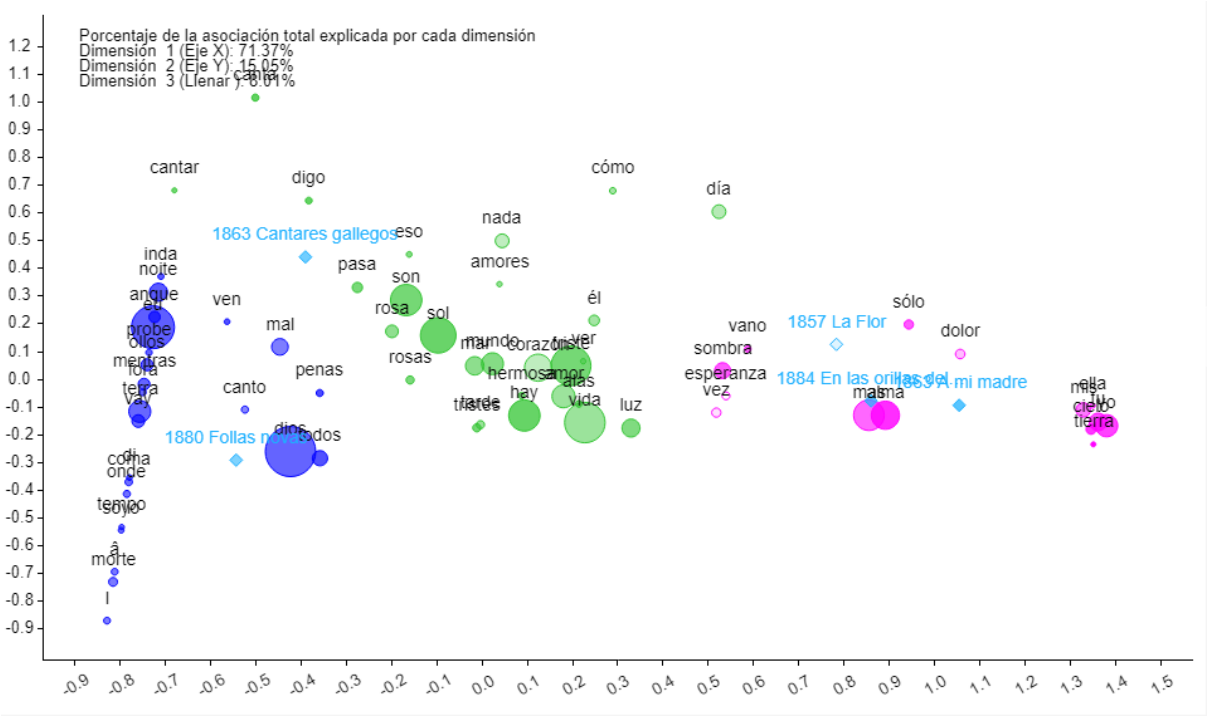

Figura 16. Análisis de correspondencias midiendo la frecuencia absoluta de los 75 términos más usados en el corpus bilingüe. Elaboración propia.

Sin embargo, si aplicamos el análisis de correspondencias midiendo la frecuencia relativa, observamos que la palabra *dios* se desplaza hacia el poema *Cantares Gallegos* (figura 17).

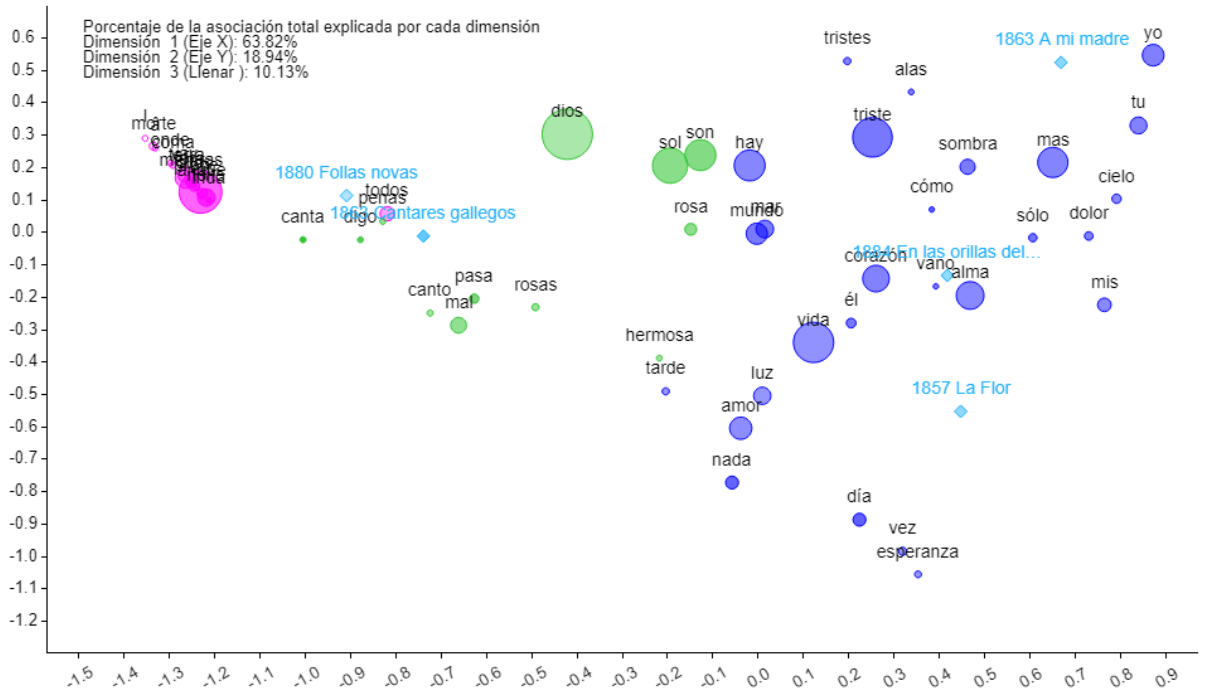

Figura 17. Análisis de correspondencias midiendo la frecuencia relativa de los 75 términos más usados en el corpus bilingüe. Elaboración propia.

Con estos diferentes métodos analíticos podremos notar que hay términos que se obtienen con una frecuencia absoluta mayor en una obra, sin embargo, no tienen gran importancia con respecto a lo que trata la obra. Si bien *dios* es una palabra que en términos absolutos está mucho más cerca de *Follas Novas* que de *Cantares Gallegos* y aún más lejano al resto de los textos, no

tiene la importancia que parecería tener cuando observamos la palabra en términos relativos y observamos que se encuentra prácticamente equidistante a ambas obras y aún más, presenta un corrimiento hacia el centro, implicando una presencia de relativa importancia en el resto de los textos.

Estos diferentes análisis podrían ser provechosos al utilizarse con otro corpus que sirviera para indagar en la similitud entre el poemario gallego de Rosalía de Castro y el de otros literarios pertenecientes al *Rexurdimento* gallego, por ejemplo. Ya que *ScatterPlot* resulta una herramienta con grandes posibilidades de investigación para este campo de trabajo.

5. ANÁLISIS ESPACIAL

Voyant Tools posee una herramienta llamada *DreamScape* que resulta provechosa para llevar a cabo análisis y visualizaciones de los espacios geográficos en la literatura, a pesar de que aún se encuentra en desarrollo. En el siguiente mapa interactivo podemos ver la ubicación cartográfica de los topónimos mencionados en los poemarios de Rosalía de Castro (figura 18).

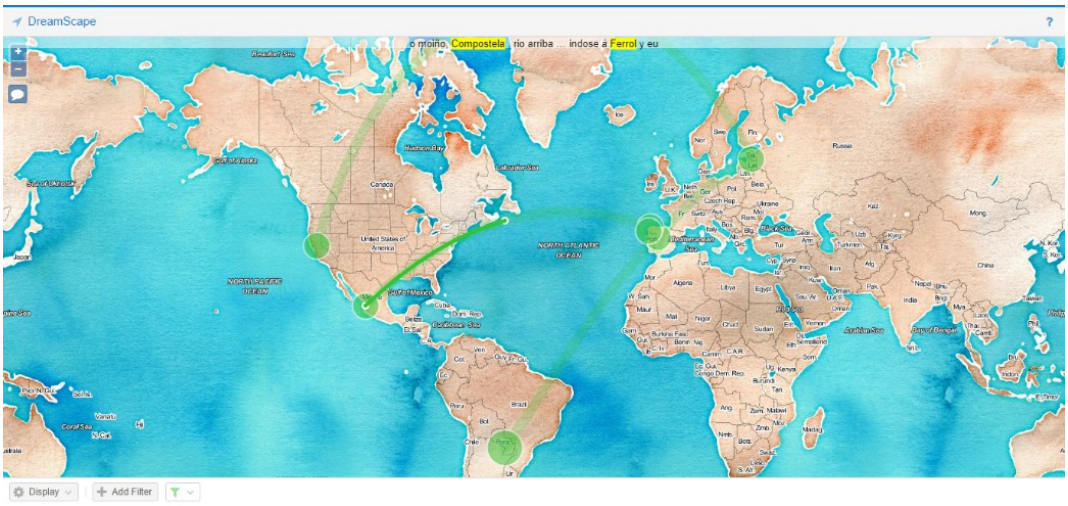

Figura 18. Primera aproximación a la herramienta DreamScape. Elaboración propia.

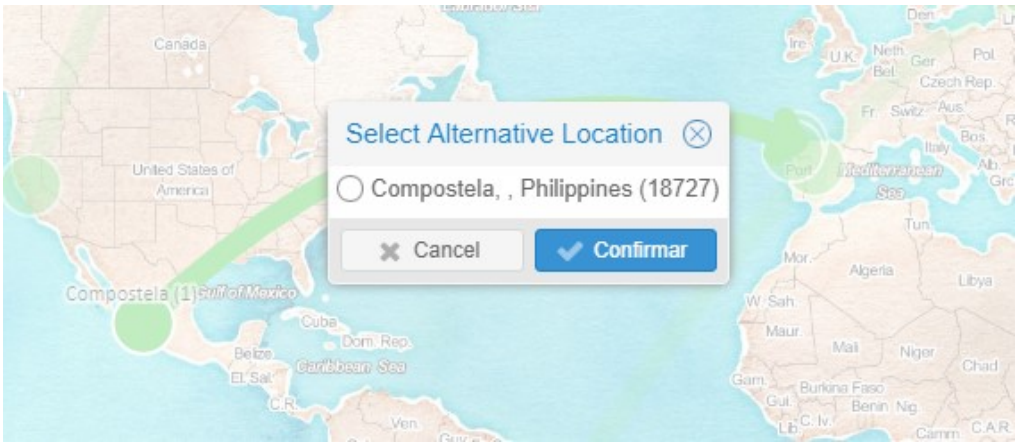

Figura 19. Intento de relocalización de la ciudad de Compostela. Elaboración propia.

Como se puede comprobar con la figura 16, la detección geográfica automática y de relocalización todavía requiere mucho trabajo de mejora, ya que no todos los topónimos han sido localizados correctamente por la herramienta. Así, por ejemplo, Voyant Tools ha marcado correctamen-

te que hay una mención a Ferrol, Cambados y Salamanca, pero en el caso de Valga y Compostela ha fallado al situarlas en Estonia y en México en lugar de en España, además que para la primera no muestra coincidencias alternativas de su ubicación y para la segunda muestra una coincidencia errónea. Siguiendo el eje de nuestro análisis, resulta interesante resaltar la fuerte presencia de escenarios de la comunidad autónoma de Galicia: Ferrol, Cambados, Compostela y Valga mientras la herramienta arroja un resultado coincidente con la Catedral de Salamanca, situada en la comunidad autónoma de Castilla y León.

6. CONCLUSIONES

Utilizando esta poderosa herramienta del campo de las tecnologías del lenguaje para analizar el corpus en cuestión, se obtuvieron resultados de lo más interesantes. Combinando las respuestas de las múltiples aplicaciones que Voyant Tools ofrece, se pudo sortear la necesidad de un estudio profundo de la obra poética de Rosalía de Castro y el contexto histórico del movimiento cultural del *Rexurdimento* gallego en el que se desarrolló su obra.

Realizando las preguntas correctas se pudo inferir a nivel micro el desarrollo de una obra en particular. Combinando las palabras claves de *Cirrus* y la relación de palabras de la herramienta *Contextos* en una línea de tendencias tenemos una visión atinada de cómo se desarrollan los versos del poema *La Flor*.

Por otro lado, al hacer una distinción de modo macro el conjunto nos cuenta de las similitudes y diferencias entre ambos lenguajes. Un rasgo diferencial importante que se destaca de su obra en gallego es la influencia del yo poético *eu*, visto con las *Líneas de Burbujas*, demostrando cierta comodidad para ser ella protagonista de su obra, a diferencia del corpus en castellano donde es más bien un recurso extraño en la autora. Una de las dificultades de este análisis es que la lectura distante nos ofrece el sentido con el que está dispuesta la palabra en un texto, por lo que debemos realizar una lectura cercana con las herramientas *Contextos* y *Lector* para obtener así el propósito de los diferentes términos que analizamos. En este sentido, las metáforas poéticas se toman en sentido literal y en términos cuantitativos, por lo que el recurso poético queda un tanto perdido y debe recuperarse también mediante una lectura cercana.

En las similitudes, se puede ver la amplia influencia religiosa que atravesó la cultura española. Demostrado en el grupo de las tres palabras más frecuentes del grupo textual encontramos a *Dios* y *vida*. Analizadas en términos absolutos en el gráfico de distancias, nos indica que *Dios* se encuentra muy presente en *Follas novas*, sin embargo, al hacer enfoque en el peso relativo de éste, denota tener gran importancia en el corpus entero, mostrándose equidistante a todos los poemarios entre sí.

Una dificultad propia de la herramienta surgió al utilizar *ScatterPlot*, tanto en la superposición de los términos en sus burbujas opacando al resto como en el color asignado de manera azarosa dificultan la correcta distinción de los términos mostrados.

Dentro de las dificultades que nos presenta este tipo de examen comparativo en un corpus bilingüe, es que los resultados cuantitativos arrojados pueden dificultar la comparación de térmi-

nos. Si bien la lengua gallega y la lengua castellana se asemejan en la escritura de un gran número de palabras, en los términos que poseen formas diferentes en su grafía dificulta el análisis ya que hay que tomar la precaución de incorporar los términos necesarios en ambas lenguas.

REFERENCIAS BIBLIOGRÁFICAS

De Castro, R. (1857). *La Flor*. Imprenta de M. González. https://bit.ly/2UBHSCM

De Castro, R. (1863). *A mi madre*. Imprenta de J. Compañel. https://bit.ly/2TNMPZ0

De Castro, R. (1863). *Cantares gallegos*. Imprenta de J. Compañel. https://bit.ly/3jVxnVj

De Castro, R. (1880). *Follas Novas*. La propaganda Literaria. https://bit.ly/3AFRNrl

De Castro, R. (1884). *En las orillas del Sar*. Establecimiento Tipográfico de Ricardo Fe. https://bit.ly/2UyNDBZ