

Big data, criticometría y el estudio de las literaturas nacionales en la bibliografía crítica: el caso excepcional de la literatura cubana

Big Data, Criticism, and the Study of National Literatures in Critical Bibliography: The Singular Case of Cuban Literature

Dirección

Clara Martínez
Cantón

Gimena del Río
Riande

Ernesto Priani

Secretaría

Romina De León

Francisco ESPINO
Université de Paris 8
francisco.espino.esteban@gmail.com
<https://orcid.org/0000-0002-8292-7971>

RESUMEN

Este artículo explora nuevas perspectivas de trabajo en la aplicación de métodos cuantitativos y textométricos al estudio de la presencia de las literaturas nacionales en la bibliografía crítica, especialmente de la literatura cubana. A través de la definición, cálculo y análisis de la evolución de lo que hemos denominado el Índice de Presencia de una Literatura Nacional (IPLN) y de la aplicación de herramientas digitales pretendemos abrir nuevas vías de trabajo en este campo.

ABSTRACT

This article explores new possibilities in the application of quantitative and textometric methods for the study of the presence of national literatures in critical bibliography, especially the one focused on Cuban literature. Through the definition, calculation and analysis of the evolution of what we have named the National Literature Presence Index (IPLN by its initials in Spanish) and the use of digital tools, we aim to improve the contributions in this field.

PALABRAS CLAVE

Criticometría, literaturas nacionales, métodos cuantitativos, métodos textométricos, literatura cubana.

KEYWORDS

Criticometry, National Literatures, Quantitative Methods, Textometric Methods, Cuban Literature.

RHD 5 (2020)

ISSN

2531-1786



1. INTRODUCCIÓN

El concepto de lectura distante (Moretti, 2013), basado en el análisis cuantitativo automático de grandes bases de datos, ha abierto nuevas vías de trabajo relacionadas con el uso de la informática en un campo amplio de disciplinas, que abarcan tanto la historia literaria como la historia del libro o la sociología de la literatura (Ruth-Ellen, 2013). El presente trabajo pretende abordar, bajo una perspectiva de *distant reading*, el estudio de la presencia de las literaturas nacionales en la bibliografía científica. La digitalización de una considerable cantidad de textos académicos —artículos científicos en su mayoría— publicados por investigadores de todo el mundo, nos abre la posibilidad, mediante herramientas informáticas, de medir cuantitativamente la literatura que ha sido privilegiada como objeto de estudio por la crítica literaria académica. Nuestro trabajo es una investigación empírica, un análisis cuantitativo de un orden de publicaciones científicas que, en ningún caso, podría llevarse a cabo con una lectura analógica. Nuestra hipótesis de partida es que la elección de las obras y los autores que se convierten en objeto de estudio para la crítica académica y su clasificación dentro de unas categorías temáticas concretas, están determinadas por prácticas de naturaleza social, política o económica. El objetivo de este trabajo es demostrar nuestra hipótesis e identificar y describir algunas de estas prácticas en el caso concreto del estudio de las literaturas nacionales de los nueve países de lengua española más poblados de América: México, Colombia, Argentina, Chile, Perú, Ecuador, Guatemala, Venezuela y Cuba. En la última parte de este trabajo, nos centraremos en el caso particular de la literatura cubana.

Una de las dificultades más extendidas en la aplicación de herramientas digitales en los estudios humanísticos es el acceso a los macrodatos o *big data*. Los datos masivos no suelen estar a disposición en un formato que nos permita procesar la información automáticamente. En primer lugar, vamos a confeccionar un corpus de datos de bibliografía científica a partir de la Modern Language Association (MLA) International Bibliography, la mayor base de datos de esta naturaleza del mundo. Para cada una de las referencias bibliográficas, nos centraremos en un conjunto de variables (fecha de publicación, título, literatura nacional...) que someteremos a un análisis métrico. Acotaremos dicho corpus de trabajo a nuestro objeto de estudio: la bibliografía crítica sobre literatura latinoamericana contemporánea en lengua española, con una atención particular a la literatura cubana.

Una vez definida la base de datos sobre la que trabajaremos, nos preguntaremos en qué medida los investigadores se han interesado por las literaturas nacionales de los nueve países de estudio escogidos, es decir, cuáles son las literaturas nacionales que han sido más estudiadas por la crítica académica y cómo ha evolucionado dicho interés. Para ello, nos basaremos en el concepto de criticometría, introducido por Carolina Ferrer (2011). Como referencia comparativa que nos permita evaluar si la atención de la academia a una literatura nacional es más o menos proporcionada, compararemos los datos bibliográficos con la variable población del país. Mediremos si la

proporción del número de referencias bibliográficas dedicadas a cada una de las literaturas nacionales respecto al total se corresponde con el porcentaje de su población respecto a la población total de los nueve países. Hemos creado el Índice de Presencia de una Literatura Nacional (IPLN). Este parámetro, que definiremos matemáticamente en el siguiente apartado, nos indicará, para un país concreto y en una fecha determinada, en qué medida el porcentaje de referencias bibliográficas críticas centradas en el estudio de su literatura nacional es proporcional al porcentaje que representa su población.

Por último, nos centraremos en el caso de la literatura cubana. Analizaremos cómo ha evolucionado en los últimos años su presencia en los estudios científicos y, mediante el uso de métodos textométricos, identificaremos algunas prácticas significativas que explican el desmesurado interés de la academia por su literatura.

Las bases de datos utilizadas tienen formato CSV. Todos los parámetros matemáticos (el IPLN y las distintas variables estadísticas), así como las gráficas que aparecen en este trabajo han sido calculados y elaborados por nosotros mediante funciones de Excel.

1.1. Criticometría

El análisis de la bibliografía crítica y de su evolución mediante herramientas informáticas es un campo poco explorado. Destacan las aportaciones de Carolina Ferrer, que dirige, en la Universidad de Quebec, el proyecto de investigación *Magellan: cartographie de la bibliographie critique sur la littérature hispano-américaine*¹, sobre el uso de herramientas digitales para cartografiar los estudios críticos de la literatura latinoamericana de los siglos XX y XXI. Ferrer (2017b) propone un análisis de las grandes bases de datos bibliográficos a través de una metodología que denomina *criticométrica*:

[...] nous utilisons l'approche de la criticométrie que nous avons introduite en 2011 en nous inspirant de la scientométrie. L'objectif de la scientométrie est de mesurer et d'analyser l'activité dans les sciences et la technologie. Par analogie, nous avons appelé cette approche la criticométrie, puisqu'elle a pour objectif de mesurer l'activité critique en arts, et plus particulièrement en littérature. La méthode ici proposée correspond à l'adéquation des indicateurs scientométriques à la réalité des bases de données utilisées en sciences humaines et en arts (p. 2).

La *criticometría* consiste, por analogía con la *cienciometría* (Price, 1963), en aplicar procedimientos métricos a variables propias de la crítica artística, en nuestro caso, en medir y analizar la evolución de indicadores explicativos de la bibliografía crítica. Aplicando dicha metodología Ferrer ha demostrado una sobrepresencia de ciertas literaturas nacionales en los estudios científicos: “nous constatons que les bibliographies critiques sur 14 littératures nationales cumulent plus de 82 % des publications sur l'ensemble mondial” (Ferrer, 2017b, p.18) o el lugar marginal que ocupan las escritoras como objeto de estudio, especialmente en Europa, a las que se les ha dedicado solamente el 9% de los trabajos (Ferrer, 2017a, p. 21).

¹ Accesible desde: <http://figura.uqam.ca/programme/projet-magellan-cartographie-de-la-bibliographie-critique-sur-la-litterature-hispano>.

Al ser la literatura un proceso insertado en un sistema global, social y de mercado, indicadores ajenos a la estricta naturaleza de las obras y de los canales de transmisión suponen un fuerte impacto en el estudio de la literatura y en su recepción. Gracias a la aplicación de métodos cuantitativos podemos analizar la correspondencia entre variables explicativas de la bibliografía crítica literaria con otras de distinta naturaleza –socioeconómica, política, etc.– que pueden estar correlacionadas con los indicadores bibliométricos. Ferrer (2015) ha demostrado una correspondencia entre el desarrollo de la bibliografía crítica en los países europeos y el salario bruto:

We were able to determine that there is an important correlation between the level of development of the critical bibliography about national literatures and the Gross National Income of the European countries (p. 10).

Pretendemos aportar nuevas perspectivas en el análisis criticométrico de la bibliografía del estudio de las literaturas nacionales de los nueve países de lengua española más poblados de América desde un punto de vista comparativo.

1.2. La base de datos bibliográficos de la Modern Language Association

La primera dificultad con la que nos encontramos para el estudio de bibliografía crítica es disponer de un acceso a un corpus de datos lo suficientemente extenso para que podamos aplicar una perspectiva de lectura distante y que, además, éstos estén disponibles en un formato compatible con las herramientas digitales que nos permiten aplicar métodos cuantitativos y textométricos.

Nos basaremos en la MLA International Bibliography, consultable mediante EBSCOhost². Se trata de la mayor base de datos bibliográficos sobre estudios críticos literarios en lenguas modernas del mundo. Dispone de casi tres millones de referencias que incluyen artículos, libros, revistas y otras publicaciones, desde 1888 hasta nuestros días. Está gestionada por la Modern Language Association, fundada en 1883 en Estados Unidos de Norteamérica:

The MLA International Bibliography is an essential tool for research in all aspects of modern languages and literature. Available as a searchable online database of more than 2.8 million records and constantly updated by scholars in the field, the bibliography is used by academics and teachers around the world to find and share their work. Students at all levels rely on the bibliography to lead them to the journals, books, Web sites, and other publications they need to succeed in their research projects (MLA)³.

Constituye un intento de recopilar toda la información bibliográfica crítica de las literaturas modernas con el objetivo de disponer de un buscador global especializado. En la página web podemos descargarnos la lista de fuentes indexadas y asciende a más de 13.000⁴. Aparece la información bibliográfica y, en los casos en los que está disponible, puede también descargarse el texto crítico en sí. Funciona en torno a una interfaz y un buscador convencionales. Por

² EBSCO es una plataforma de búsqueda especializada que integra varias bases de datos. El acceso no es gratuito. Puede consultarse en: <https://www.ebsco.com>.

³ Accesible desde: <https://www.mla.org/Publications/MLA-International-Bibliography/About-the-MLA-International-Bibliography>.

⁴ La lista completa puede ser descargada en <https://www.mla.org/Publications/MLA-International-Bibliography/Frequently-Asked-Questions> bajo el epígrafe “How many journals are covered in the bibliography?”.

ejemplo, si queremos buscar información bibliográfica de Borges, basta con hacer una búsqueda con su nombre y nos aparecen unas cinco mil referencias bibliográficas cuyo autor principal de estudio es Borges. Si accedemos a cualquiera de las referencias, la interfaz nos proporciona la información explicativa convencional de una cita bibliográfica, como el título, el autor o la fecha. Pero, además, la base de datos consta de un sistema de etiquetas en formato *Thesaurus*, que permite precisar la búsqueda bibliográfica:

MLA Bibliography Thesaurus is a searchable and browsable controlled vocabulary that enables users to identify terminology appearing in current scholarship and to create more precise searches (MLA)⁵.

El sistema le atribuye a cada referencia bibliográfica un conjunto de etiquetas que nos facilita llevar a cabo búsquedas detalladas. Son heterogéneas, pueden referirse, tanto en inglés como en español, a las literaturas nacionales —*Cuban literature, Argentine literature*—, a categorías críticas preestablecidas de diversa índole —*psicoanálisis, autobiografía, novela histórica*— o a términos más abstractos, generales y/o ambiguos —*exilio, violencia, Barroco*—. La interfaz también nos permite filtrar los resultados según algunos parámetros como fecha o tipo de texto. Podríamos consultar, por ejemplo, la información bibliográfica de todas las referencias de revistas universitarias cuyo principal autor de estudio es Borges, que de alguna forma problematizan su obra con la dictadura argentina y que han sido publicadas en los años noventa. Es una base de datos indexada muy rica pero convencional.

Lo interesante para nuestro trabajo es la posibilidad de descargar los datos de las búsquedas, en los que se incluyen todas estas variables, en un formato informático. Podemos recuperar dichos datos con la extensión Research Information System (RIS), compatible con gestores de referencias bibliográficas como Zotero, pero también en formato CSV o XML, lo que nos facilita el uso de una plantilla Excel y de los principales lenguajes de programación usados en el tratamiento automático de la lengua (TAL), R y Python.

En definitiva, si tenemos posibilidad de pago, disponemos de la mayor base de datos del mundo sobre referencias bibliográficas de las literaturas en lenguas modernas en formatos adecuados para ser manipulados con herramientas digitales. En este trabajo vamos a aplicar métodos cuantitativos y textométricos a este corpus de datos. Nos centraremos en el análisis de las referencias etiquetadas como literaturas nacionales de los nueve países que hemos elegido como objeto de estudio: las referencias bibliográficas catalogadas como literatura mexicana, colombiana, argentina, peruana, chilena, guatemalteca, ecuatoriana, venezolana o cubana. Se trata de más de setenta mil muestras y cada una de ellas está compuesta por decenas de variables.

⁵ Accesible desde: <https://www.mla.org/Publications/MLA-International-Bibliography/About-the-MLA-International-Bibliography/MLA-Thesaurus>.

2. UNA FOTOGRAFÍA CUANTITATIVA DE LA PRESENCIA DE LAS LITERATURAS NACIONALES EN LA BIBLIOGRAFÍA CRÍTICA

¿En qué porcentaje las literaturas nacionales están actualmente representadas como objeto de estudio en la bibliografía crítica? ¿Se distribuyen regularmente o destaca algún interés crítico particular en alguna de las nueve literaturas nacionales que estamos estudiando? Vamos a realizar una fotografía cuantitativa de nuestra base de datos bibliográficos de las referencias que han sido catalogadas como dedicadas al estudio de alguna literatura nacional. La interfaz de la MLA International Bibliography nos permite acceder a los datos filtrando por esa etiqueta y un periodo temporal elegido. Hemos recopilado en formato CSV el número de referencias bibliográficas dedicadas a cada una de las literaturas nacionales publicadas entre 1888, que es la primera fecha en la que se tienen datos, y diciembre de 2018. Así, por ejemplo, aparecen del orden de 10.000 referencias bibliográficas que han sido catalogadas por la MLA como referidas a literatura cubana publicadas en ese periodo, lo que supone un 14% del total de referencias bibliográficas dedicadas al conjunto de las literaturas nacionales que estamos estudiando. Hemos calculado la distribución de las nueve literaturas y el resultado es el siguiente:



Figura 1. Distribución de las referencias bibliográficas críticas centradas en las literaturas nacionales publicadas entre 1888 y 2018 de los nueve países de habla española más poblados de América. Elaboración propia.

Vemos (figura 1) que la distribución por literaturas nacionales es bastante desigual. Las referencias bibliográficas centradas en el estudio de las literaturas mexicana y argentina acumuladas durante estos ciento treinta años (1888-2018) son más significativas que las focalizadas en las de otros países. Juntas superan el cincuenta por ciento del total de referencias de las nueve literaturas nacionales, es decir que el espacio que la crítica le ha dedicado a estas dos literaturas es mucho más significativo que al resto. En tercer lugar, aparece la literatura cubana. En el otro extremo, algunas literaturas nacionales han recibido una escasísima atención crítica, como las literaturas

de Ecuador y de Guatemala, que apenas cuentan con el tres por ciento del total de las referencias bibliográficas.

3. UN ANÁLISIS TEMPORAL Y RELATIVO A LA POBLACIÓN. ÍNDICE DE PRESENCIA DE UNA LITERATURA NACIONAL (IPLN)

Los datos anteriores muestran la cantidad de bibliografía crítica dedicada a cada una de las literaturas en valor absoluto, pero necesitamos relacionarlos con otras variables que nos permitan medir el peso que cada país tiene en la creación y publicación de las obras literarias. Las posibilidades son varias, podríamos pensar en la cantidad de libros publicados, el número de editoriales o de escritores y escritoras. Cualquiera de estas opciones, aunque imperfecta, podría abrirnos interesantes vías de trabajo. Debido al espacio reducido al que debemos ceñirnos, vamos a medir la proporcionalidad respecto a la población de cada país. Aunque la distribución demográfica no será la única variable que explique la cantidad y el valor de la creación literaria de un país, debería haber una relación directa entre la cantidad de población y su *potencial* para crear y publicar obras que fueran de interés para la crítica científica. Desde un punto de vista cuantitativo, un país con mucha más población que otro, debería producir más obras literarias que resultasen de interés para la comunidad científica. Nuestro objetivo es detectar si existen grandes anomalías en estas proporciones que nos abran nuevas perspectivas en el estudio de las relaciones entre la política, la sociedad, la historia y la presencia de las literaturas nacionales en la bibliografía crítica.

Para medir la evolución de la población en los nueve países utilizaremos la base de datos del Banco Mundial⁶, quien pone a disposición, de forma gratuita y en línea, series cronológicas de diversos índices socioeconómicos, que van desde la demografía a indicadores de desarrollo, macroeconómicos, de educación o salud, en formato CSV o XML, para cualquier país. Hemos recuperado la información de la variable *población total* de los nueve países que son objeto de nuestro estudio, detallada por año, de 1960, primer año del que se dispone de datos, a 2018.

Para una literatura nacional X y para un año dado Y, definimos el Índice de Presencia de una Literatura Nacional (en adelante, IPLN de X en Y) como:

$$= \frac{\text{ref.biblio. cuyo objeto de estudio es la literatura X acumuladas hasta Y} / \text{Ref. biblio. totales acumuladas hasta Y}}{\text{Poblacion de X en Y} / \text{Población total en Y}}$$

Pretendemos de esta forma medir la presencia que el estudio de una literatura nacional tiene en los trabajos críticos en relación con su población. El IPLN matemáticamente expresa el grado de sobrerrepresentación o subrepresentación del número de las referencias bibliográficas cuyo objeto de estudio es una literatura nacional respecto al porcentaje de su población. Precisa de dos entradas, el país y el año. Tal y como lo hemos definido, los valores límites oscilan entre 0 e infinito. Un valor entre 0 y 1 se corresponderá con una presencia de su literatura nacional en la biblio-

⁶ Puede consultarse en: <https://datos.bancomundial.org/>.

grafía crítica inferior a la representación de su población. Si es igual a 1 se tratará de una presencia exactamente proporcional a la de su población, mientras que un índice mayor que la unidad evidenciará una sobrerrepresentación. Así, por ejemplo, el IPLN de Argentina en 1960, se corresponderá con el cociente entre el porcentaje de referencias bibliográficas relativas a la literatura argentina recogidas en la base de datos de la MLA International Bibliography desde 1888 hasta 1960 respecto al total de referencias bibliográficas centradas en las literaturas de los nueve países hasta ese mismo año, y entre el porcentaje de su población respecto a la población total de los nueve países en esa fecha. Si su valor fuera 2 significaría que la presencia de su literatura en la crítica científica está sobrerrepresentada respecto a su población, en un orden del doble. Si fuera 0,2 significaría que está subrepresentada, en un orden de $1/5$. Hemos calculado el IPLN en los nueve países, con fecha diciembre de 2018 y el resultado es el siguiente:

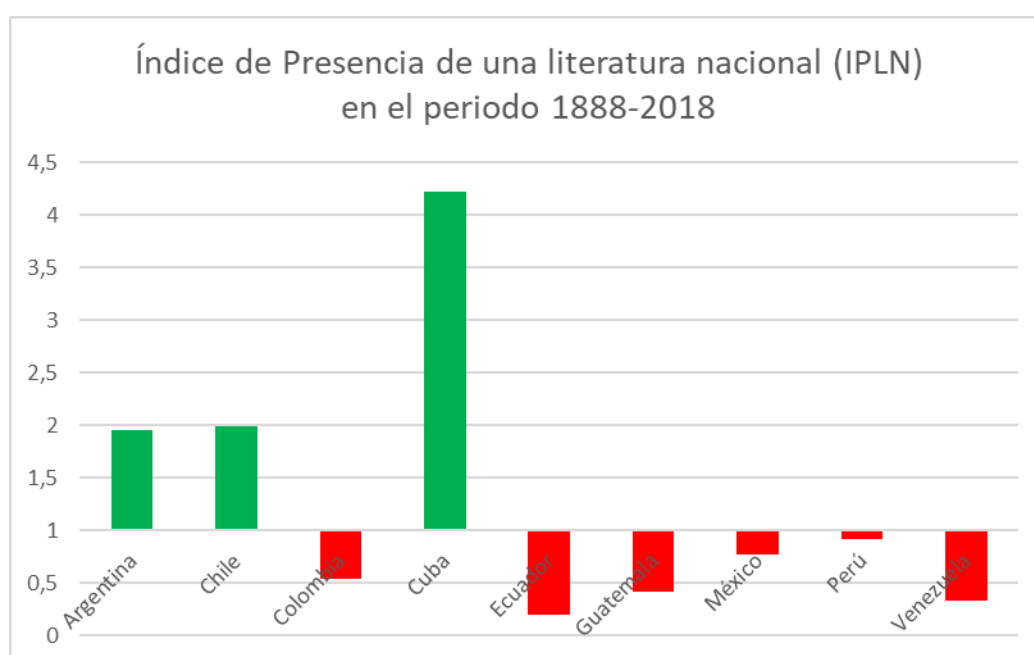


Figura 2. IPLN de los nuevos países estudiados en 2018. Elaboración propia.

La figura 2 representa, por tanto, la presencia de las nueve literaturas nacionales en la bibliografía crítica respecto a su población en el periodo 1888-2018. Destaca una variada disimilitud. Las literaturas argentina, chilena y cubana están sobrerrepresentadas respecto a su población —un IPLN mayor que la unidad—. El número de referencias bibliográficas publicadas hasta diciembre de 2018 es proporcionalmente mayor que la población que representan. El resto de las literaturas, por el contrario, están subrepresentadas —un IPLN menor que la unidad—. Las literaturas venezolanas y ecuatorianas son las más subrepresentadas de las nueve. El caso de Cuba, en el que nos centraremos posteriormente, es muy significativo y no tiene comparación con el de ninguna otra literatura nacional. Hemos calculado el consciente entre el IPLN de Cuba y del resto de países (figura 3). El IPLN de Cuba en 2018 es el doble del de las otras dos literaturas sobrerrepresentadas, la argentina y la chilena, del orden de más de cinco veces el de la literatura mexicana, ocho

veces el de la colombiana, trece veces el de la literatura venezolana y más de veinte veces el de la literatura ecuatoriana (figura 3).

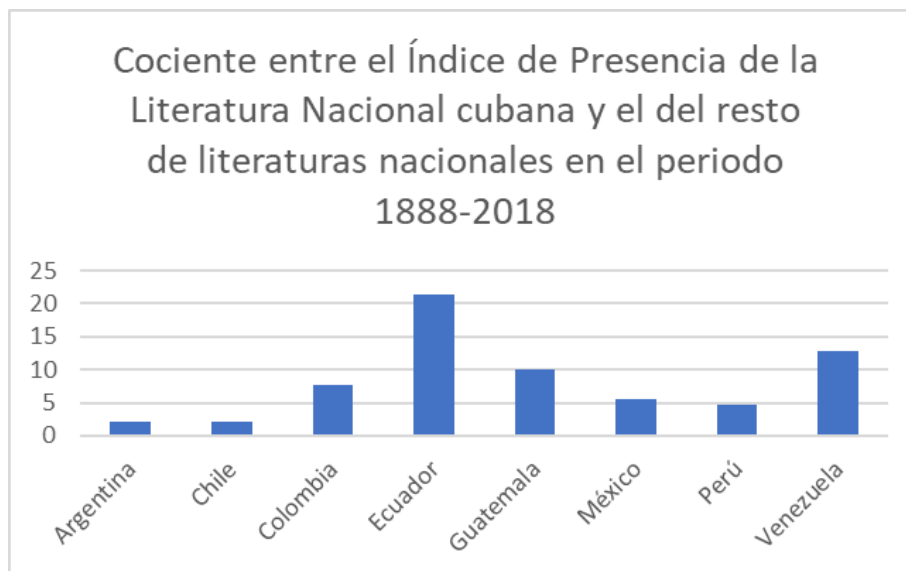


Figura 3. Cociente entre el IPLN de Cuba en 2018 y el del resto de literaturas nacionales. Elaboración propia.

Si el IPLN demuestra que no hay relación alguna entre la cantidad de población de Cuba y la presencia de su literatura en la bibliografía crítica, es decir, que el motivo por el que se ha estudiado más la literatura cubana no está relacionado con su demografía y, por tanto, con un mayor potencial de escritores y obras, ¿qué factores sociopolíticos, económicos o de otro tipo, explican esta disparidad en el interés que su literatura ha despertado? Las reflexiones sobre la construcción del canon literario y las narrativas nacionales han sido plurales. Para Iván de la Nuez (1998), desde el comienzo de las independencias en el siglo XIX, el canon latinoamericano se ha construido buscando un sujeto que desmarque a las nuevas naciones de los Estados Unidos de Norteamérica y de Europa (pp. 103-116), prevalecen los autores y los textos donde mejor se narra una nación diferente de los dos modelos occidentales. Para Bhabha (1994), las naciones pasan de ser sujetos pasivos descritos en los textos a *narrativas narradoras*, a inventar sus propias genealogías a través de la escritura (pp. 291-322). Ahora bien, ¿cómo ha operado la presencia en la bibliografía crítica en esa relación compleja entre nación, política y literatura nacional, por qué aparecen tales desproporciones entre los países latinoamericanos respecto a su población y por qué es Cuba un caso excepcional? La respuesta a estas diferencias abismales es, probablemente, múltiple y compleja, pero el análisis del IPLN demuestra el interés en avanzar en la aplicación de métodos cuantitativos. Vamos a profundizar en el análisis de la base de datos para intentar encontrar prácticas que respondan a estas preguntas.

Al tratarse de un índice que considera la bibliografía crítica acumulada según una fecha de entrada, es decir, tenemos en cuenta toda la bibliografía crítica publicada entre 1888 –primer año del que disponemos de datos– y esa fecha, vamos a medir si las disparidades han sido uniformes en las últimas décadas o si han ido evolucionando según la variable temporal. Para ello, hemos calculado el IPLN de los nueve países en las siete décadas anteriores: 1960, 1970, 1980,

1990, 2000, 2010 y 2018. El primer caso consideraría la cantidad de bibliografía crítica publicada entre 1888 y 1960, y la población de 1960; el segundo entre 1888 y 1970, y la población de 1970; y así sucesivamente hasta 2018. El resultado es el siguiente⁷:

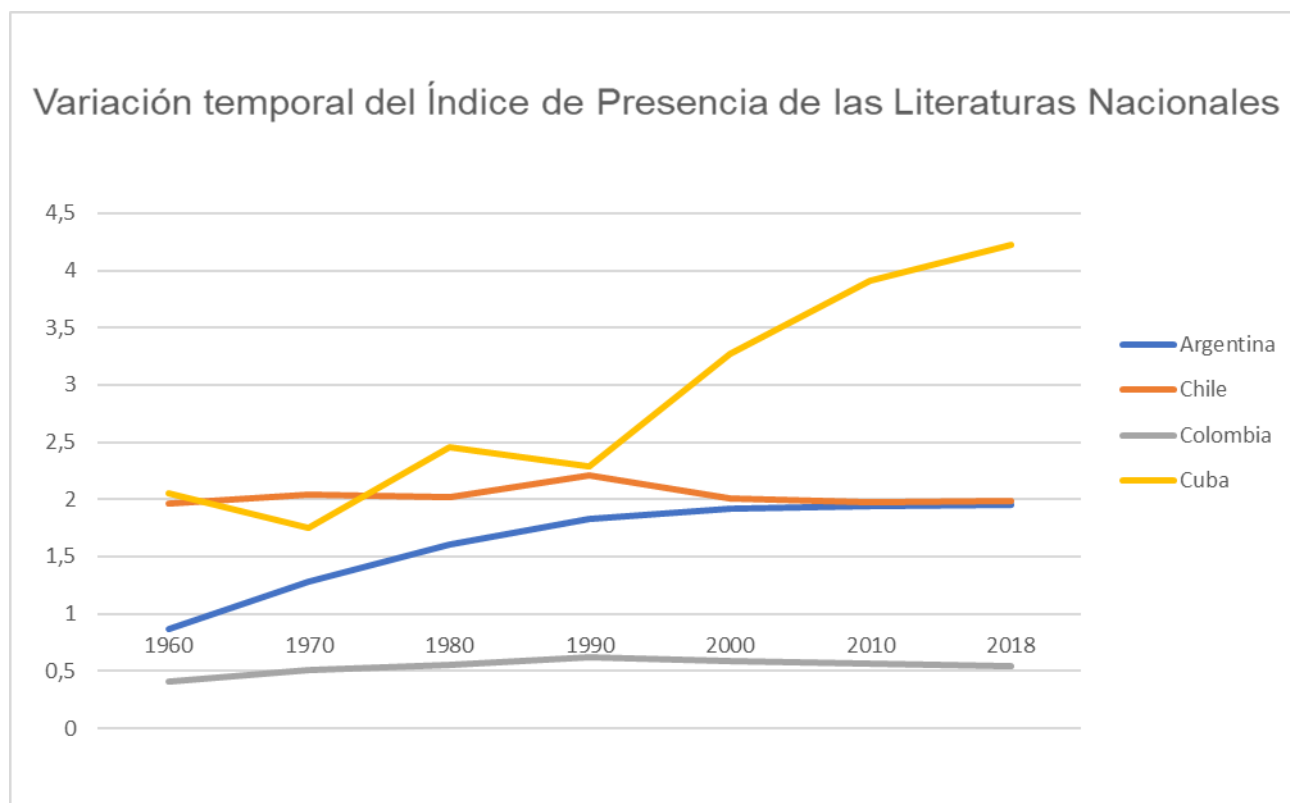


Figura 4. Variación temporal del IPLN de las literaturas argentina, chilena, colombiana y cubana. Elaboración propia.

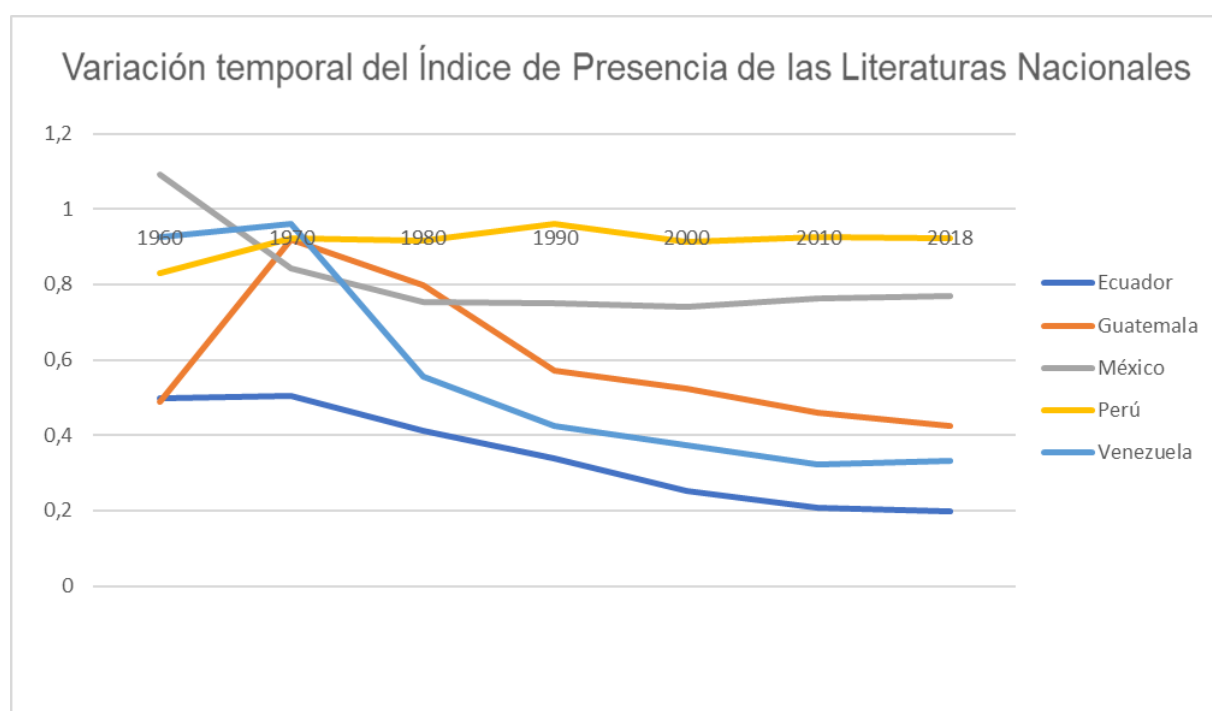


Figura 5. Variación temporal del IPLN de las literaturas ecuatoriana, guatemalteca, mexicana, chilena, peruana y venezolana. Elaboración propia.

⁷ La lectura de los nueve países en una única gráfica resulta farragosa, así que hemos separado el resultado en dos gráficas, que son en realidad el mismo análisis.

La casuística de los nueve países (figuras 4 y 5) es casi tan plural como el número de literaturas nacionales. Las diferencias entre los países son muy significativas en cualquier periodo analizado. En primer lugar, tenemos un grupo de literaturas cuya presencia ha sido muy inferior a la proporción de su población en los últimos sesenta años. Es el caso de las literaturas colombiana, ecuatoriana, venezolana y guatemalteca. La literatura con una subrepresentación más acusada es la ecuatoriana. En los años sesenta y setenta la subrepresentación era del orden de la mitad, y ha ido disminuyendo hasta situarse en un IPLN de 0,2, es decir, cinco veces menor que la representación de su población. De igual forma, en los años sesenta la presencia de la literatura colombiana se correspondía con un IPLN de 0,5, pero esta se ha mantenido en una subrepresentación estable, sin grandes variaciones hasta nuestros días. Durante los años sesenta la presencia de la literatura venezolana en la bibliografía crítica era casi del mismo orden que su población, pero después sufre una fortísima bajada exponencial hasta situarse, en 2018, con un IPLN de 0,3. Igualmente le ocurre a la literatura guatemalteca que aumenta su presencia considerablemente de los años sesenta a setenta, hasta colocarse prácticamente a un nivel proporcional a su población, pero luego sufre una caída lineal interrumpida hasta un IPLN de 0,4 en 2018. La literatura mexicana es la única de las literaturas nacionales que evoluciona de una sobrerrepresentación en los años sesenta a una subrepresentación del orden del 0,8 en 2018. En la presencia de la literatura peruana también destaca una particularidad excepcional: se mantiene estable en un orden casi proporcional a su población ininterrumpidamente desde los años cincuenta hasta hoy día.

En el lado opuesto, observamos que hay tres literaturas nacionales que están sobrerrepresentadas: la argentina, la chilena y la cubana. La literatura chilena se ha mantenido sin apenas variaciones con un IPLN del orden de 2 en los últimos sesenta años. La presencia de la literatura argentina ha evolucionado de una mínima subrepresentación en los años sesenta a una sobrerrepresentación del orden de 2 en 2018, al igual que la chilena.

Mediante el uso de métodos cuantitativos y la definición y posterior cálculo del IPLN hemos demostrado una distribución muy desigual de la presencia de las literaturas nacionales en la bibliografía crítica respecto a su población. Las literaturas nacionales no han despertado en la crítica científica un interés proporcional a una lógica poblacional y estas diferencias se manifiestan en los últimos sesenta años. Se ha privilegiado el estudio de las literaturas cubana, argentina y chilena y se le ha prestado particularmente poca atención a las literaturas venezolanas y ecuatorianas. ¿Por qué? Entender las razones que explican estos fenómenos y su evolución en el tiempo requiere un estudio profundo de las miles de referencias bibliográficas de nuestra base de datos que sobrepasa el espacio de este artículo. Las conclusiones de este trabajo pretenden abrir las puertas a futuros estudios cuantitativos que puedan desentrañar las prácticas ligadas a los fenómenos aquí demostrados. Vamos a profundizar en el caso más significativo, el de la literatura cubana.

3. EL CASO PARTICULAR DE LA LITERATURA CUBANA

El caso de la presencia de la literatura cubana en la bibliografía crítica es particularmente excepcional. El estudio de su literatura ha estado mucho más presente en los trabajos críticos que cualquier otra literatura nacional con relación a su población, en todas las épocas desde que disponemos de datos. Hemos calculado la evolución en el tiempo de la desproporción del IPLN de la literatura cubana respecto al valor medio del IPLN de las literaturas del resto de países estudiados y el resultado es el siguiente:

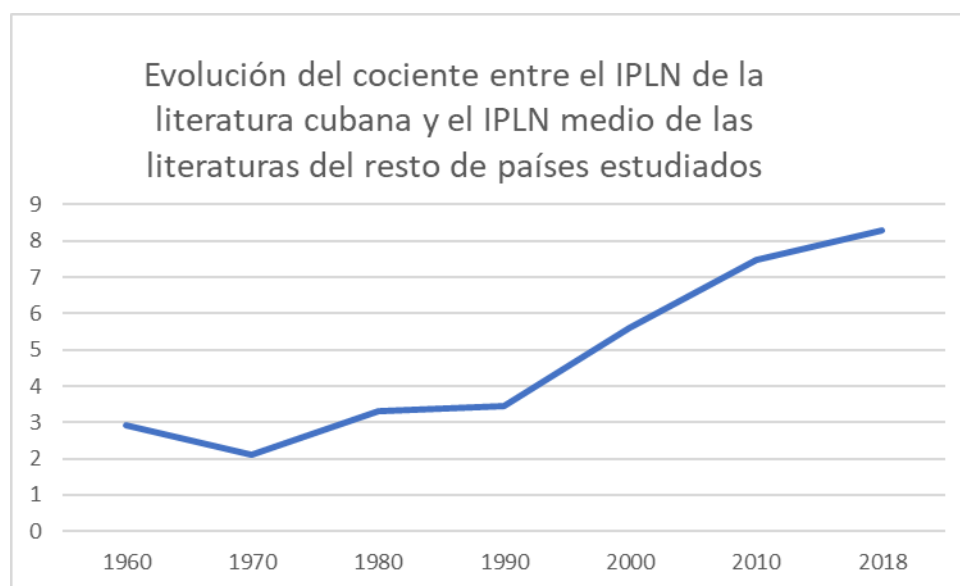


Figura 6. Evolución temporal del cociente entre el IPLN de la literatura cubana y el IPLN medio de las literaturas nacionales que son objeto de estudio en este trabajo. Elaboración propia.

En todo momento, el IPLN de la literatura cubana ha sido muy superior al IPLN medio del total de los países (figura 6). El cociente ha variado entre un valor de 3, en 1960, y algo más de 8, en 2018. Es decir, en 1960 la presencia de la literatura cubana en la bibliografía crítica era el triple que la presencia media de las literaturas de los nueve países, y este valor ha ascendido al orden de 8 veces en 2018. Vemos que es especialmente a partir de los años 90 (figura 6) cuando se produce la mayor fase de crecimiento. ¿Por qué la presencia de la literatura cubana ha acaparado un interés crítico tan desproporcionado? (Rojas, 2000, p. 66). sugiere que la sobrepresencia de la literatura cubana en el influente canon occidental defendido por Bloom (1995) —en el que los escritores cubanos representan un tercio del total de los latinoamericanos incluidos en el canon—, es debida a la cercanía entre el autor y el profesor cubano de la Universidad de Yale Roberto González Echevarría. Podríamos preguntarnos si es el interés crítico lo que delimita el canon, o si es la canonización de los autores lo que implica el interés académico por su obra. En cualquier caso, sólo a través de métodos cuantitativos conseguiremos descifrar los fenómenos que aclaren ésta y otras muchas disparidades, puesto que disponemos de 10820 referencias bibliográficas dedicadas a la literatura cubana publicadas hasta diciembre de 2018. Vamos a profundizar en el estudio de esta base de datos.

La distribución anual de las referencias bibliográficas es la siguiente:

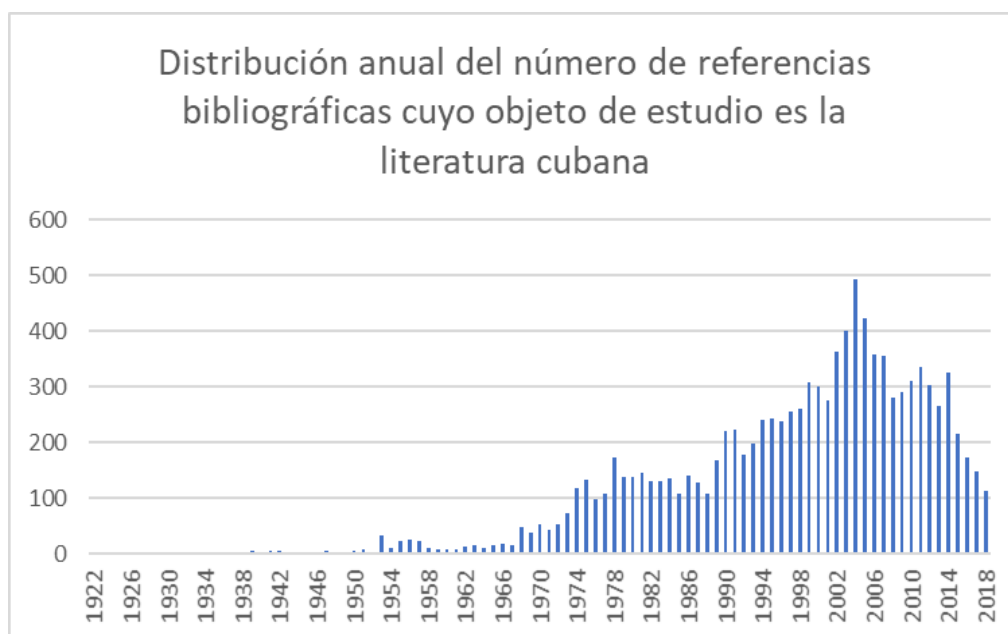


Figura 7. Distribución anual de las referencias bibliográficas sobre literatura cubana. Elaboración propia.

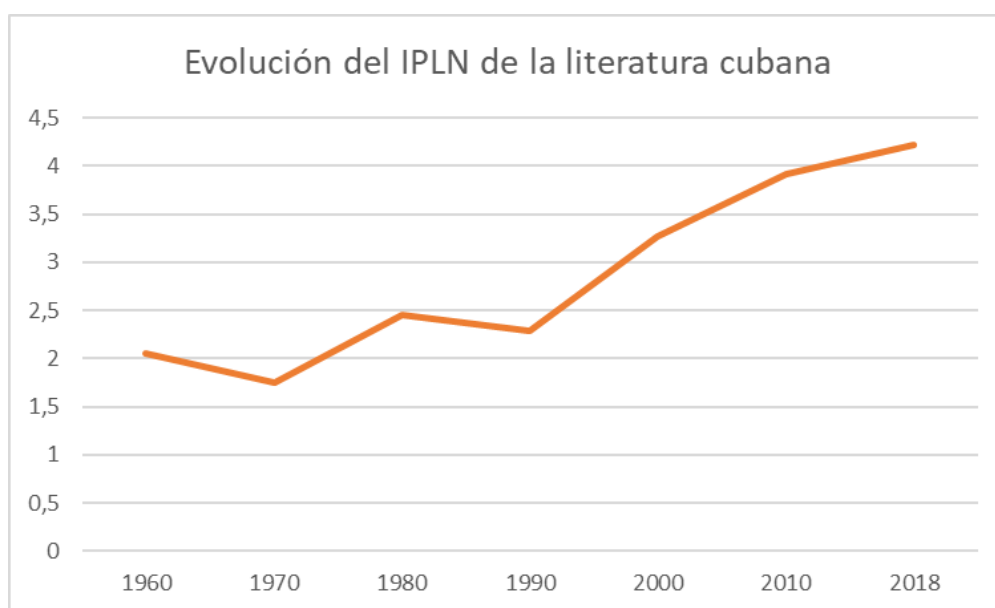


Figura 8. Evolución temporal del IPLN de la literatura cubana. Elaboración propia.

Si comparamos el número de referencias bibliográfica anuales (figura 7) con la evolución temporal del IPLN de Cuba (figura 8), observamos que las dos fases de crecimiento más significativas, inauguradas a finales de los años sesenta y a principio de los noventa, coinciden. La tasa de crecimiento de la población de los nueve países que son objeto de estudio ha sido bastante similar desde los años sesenta, luego parece lógico que el incremento de la presencia de la literatura cubana en la bibliografía crítica venga determinado por un aumento del número de referencias en valor absoluto. Vamos a dividir nuestro análisis en tres periodos temporales limitados por estos dos puntos de inflexión en el número de referencias bibliográficas: hasta 1968, entre 1968 y 1989 y a partir del año 90.

Una vez aplicada una división basada en la correlación con la variable temporal, realizaremos un análisis textométrico. La textometría es una disciplina desarrollada esencialmente en Francia a partir de los años 70 basada en la estadística textual. Consiste en desarrollar modelos estadísticos en datos textuales como la repetición de palabras, sintagmas, indicadores de evolución léxica o detección de rupturas significativas (Heiden y Pincemin, 2008). Para cada referencia bibliográfica disponemos de más de 50 variables. Nos encontramos con un sistema multivariable estructurado demasiado extenso. Vamos a limitar la matriz aplicando un método factorial, es decir, reduciremos la base de datos a las variables que pueden estar correladas entre sí. En una primera aproximación vamos a aislar la matriz con las variables *title*, el título de la referencia bibliográfica, y *tags*, las etiquetas atribuidas por la MLA International Bibliography. Aparecen, por ejemplo, referencias a las que se les ha asignado la etiqueta *literatura 1800-1899* o *psicoanálisis*.

En el análisis de la evolución temporal la variable *tags* puede darnos una información explícita sobre las características de la bibliografía crítica, ya que está especialmente concebida como una herramienta facilitadora de la búsqueda y explicación de la referencia. Respecto al título, ¿qué información puede revelarnos en una referencia académica? En el artículo *Reflections on Seven Thousand Titles (British Novels, 1740-1850)*, Moretti (2009) analiza siete mil títulos de novelas inglesas publicadas entre 1740 y 1850 y se pregunta sobre la pertinencia de analizar los títulos. Pensando en términos globales, el título es la codificación de un mensaje en un contexto de mercado (p. 134), una elección consciente y central, una decisión deliberada. Además, por definición, es la parte más pública del libro (p. 145). El título es un elemento esencial de la creación, la publicación, la presencia y la recepción de la obra. Si nos focalizamos en una referencia bibliográfica académica, la importancia y el interés exhibicionista del título se multiplican exponencialmente respecto a una novela. La propia naturaleza divulgadora de la investigación impone, en la mayoría de los casos, que el título opere como un resumen o como emisor de los conceptos académicos claves del trabajo. La academia es un sistema de circulación y búsqueda semántico: si el trabajo se basa en una categoría crítica concreta y/o un corpus específico, con bastante probabilidad aparecerá en el título. El propio procedimiento de consumo bibliográfico habitual, la búsqueda en catálogos suele basarse en la información del título.

Para realizar el análisis textométrico de estas dos variables en las diez mil referencias relativas a la literatura cubana, debemos preparar del corpus de datos. Se han desarrollado numerosos lenguajes de programación y herramientas informáticas que nos facilitan la aplicación de análisis textométricos a un texto. Sólo nos interesa la parte que contiene información semántica y ésta ha de estar estructurada en un formato adecuado. Así, al texto que hemos recuperado en extensión CSV, le aplicaremos los procesos de tokenización, lematización y eliminación de palabras vacías usando la librería *spaCy* de Python⁹. La tokenización transforma la secuencia de caracteres

⁹ *SpaCy* es una librería de Python especialmente diseñada para el tratamiento automático de la lengua (TAL). Las especificaciones técnicas pueden consultarse en: <https://spacy.io/>. Hemos utilizados los APIs específicos dedicados a la tokenización: <https://spacy.io/api/tokenizer> y a la lematización: <https://spacy.io/api/lemmatizer>.

en una de tokens, que son elementos con significado. La lematización consiste en la reducción de cada palabra a su lema, la forma de base que encontraríamos en un diccionario (Schoch, 2017). Disponemos así de más de diez mil referencias con las variables *title* y *tags*, debidamente procesadas para aplicarle un análisis textométrico. Utilizaremos una combinación de la herramienta desarrollada por el École Normale de Lyon¹⁰, basada en una plataforma TXM (Heiden, Magué y Pince-min, 2010) de libre acceso y gratuita, que incorpora un análisis de estadística textual completo mediante una interfaz sencilla y las funciones de Excel.

El resultado del análisis textométrico de los más de diez mil referencias de las que disponemos en nuestro corpus de datos sobre literatura cubana arrojan tal cantidad de perspectivas de trabajo que recorrerlas en detalle sobrepasaría con creces el espacio del que disponemos. Lo que pretendemos es explicar algunos de los fenómenos que hemos identificado y que dan una respuesta parcial al estudio de la evolución y la disparidad de la presencia de la literatura cubana en la bibliografía crítica. Esbozaremos una cartografía de autores y nos centraremos en cada uno de los periodos de estudio.

3.1. Cartografía de los autores cubanos estudiados en la bibliografía crítica

En una primera lectura de los resultados de los análisis textométricos, hemos identificado una fuerte presencia de nombres propios de escritores en el título de las referencias bibliográficas. Vamos a realizar una cartografía de autores, es decir, una fotografía cuantitativa de los principales escritores que son objeto de estudio en la literatura cubana. Para ello, volvemos a la interfaz de la MLA, que nos permite una búsqueda basándonos en la etiqueta *primary subject author*. Hemos elegido a José Martí, el escritor más citado del siglo XIX y a los seis escritores más representados del siglo XX, que son, además, los seis cubanos incluidos en el canon occidental de Bloom (1995): Alejo Carpentier, Lezama Lima, Guillermo Cabrera Infante, Reinaldo Arenas, Severo Sarduy y Nicolás Guillén. El resultado de las referencias bibliográficas que tiene como principal objeto de estudios estos escritores es el siguiente:

¹⁰ Accesible desde: <http://textometrie.ens-lyon.fr/>. Basta con alimentar la plataforma con un corpus de datos específico para obtener un análisis textométrico del mismo. Puede consultarse un manual de uso detallado: <http://textometrie.ens-lyon.fr/files/documentation/Manuel%20de%20TXM%200.7%20FR.pdf>.

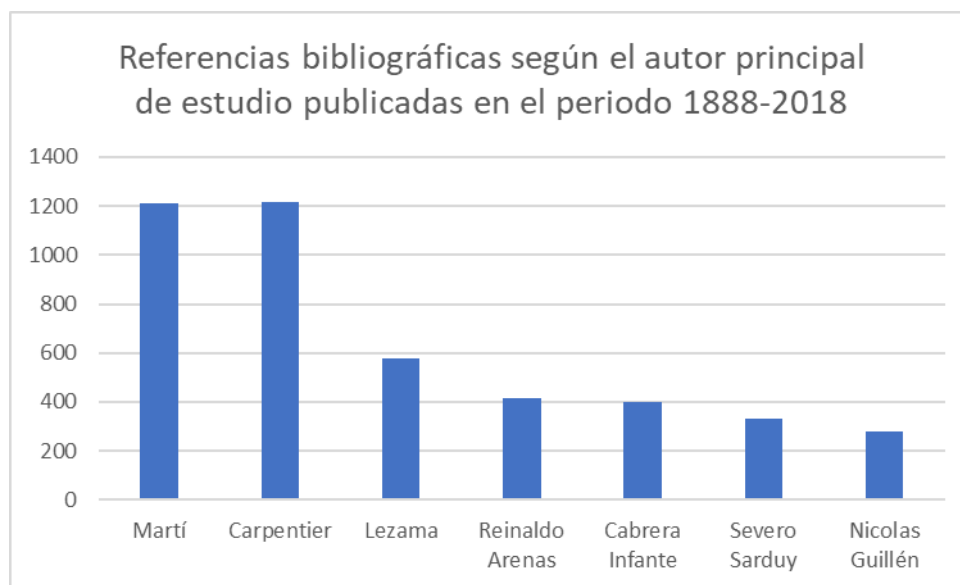


Figura 9. Referencias bibliográficas acumuladas en 2018 según el autor principal de estudio. Elaboración propia.

Destaca (figura 9) una fuerte presencia de Martí y Carpentier, los únicos escritores cubanos que suman más de mil referencias bibliográficas. Algo más de la mitad están dedicadas a Lezama Lima. En un tercer nivel aparecen Reinaldo Arenas y Cabrera Infante, del orden de cuatrocientas. Por último, Severo Sarduy y Nicolás Guillén, acumulan ambos unas trescientas. Sea cual sea el criterio que haya prevalecido en la construcción del canon cubano, bien la percepción común de cómo el texto contribuye a forjar la identidad nacional como sugiere el crítico Santí (1996, p. 369), bien cualquier otro, la combinación de análisis textométricos y cuantitativos a las referencias bibliográficas pueden arrojar mucha luz a su estudio.

3.2. La presencia de la literatura cubana en la bibliografía crítica en el período 1888-1968

La presencia de la literatura cubana hasta 1968 es especialmente alta comparada con el resto de las literaturas nacionales, del orden de tres veces el IPLN medio (figura 6). Se sitúa en un orden similar al de la literatura chilena y muy por encima del resto de literaturas nacionales analizadas (figuras 4 y 5). Hemos aplicado, según la metodología anteriormente descrita, un análisis textométrico a las variables *title* y *tags* de las referencias publicadas desde que aparecen datos hasta esa fecha. En el resultado de la estadística textual observamos que destaca la notoria representación de dos términos: *Martí* y *1800-1899*. La mayoría de las referencias bibliográficas son de los años cincuenta y sesenta. A pesar de haber transcurrido la mitad del siglo, el interés crítico está muy centralizado en la literatura del siglo XIX, bajo la etiqueta *1800-1899*. El 67% de las referencias están centradas en la literatura decimonónica. Si lo comparamos con el resto de los periodos, las referencias que tienen como objeto de estudio la literatura del siglo XIX desciende al orden del 22%, tanto entre 1968 y 1980 como tras 1990.

Este interés por la literatura del siglo XIX viene determinado en gran medida por la figura de Martí. La presencia de una literatura nacional puede estar fuertemente condicionada por lo que llamaremos un efecto de autor: un interés crítico desproporcionado en un autor concreto asociado a

una literatura nacional que acapara en solitario la mayoría de la presencia literaria en los estudios científicos. En un 38% del total de los títulos de las referencias de esta etapa aparece explícitamente una alusión a Martí, una cifra mucho mayor que la de otros los autores estudiados de la misma época, como Julián del Casal, Cirilo Villaverde o Gertrudis Gómez de Avellaneda. Los títulos que aluden directamente a estos tres últimos autores no suman juntos ni un 10% del total. Este efecto de autor ha reducido considerablemente el interés crítico por sus contemporáneos.

Otro aspecto que tenemos que destacar es la escasa presencia e incluso ausencia de los grandes nombres de la primera mitad de siglo. Lezama Lima y Virgilio Piñera están prácticamente ausentes. En el caso de Lezama, *Paradiso*, su obra magna, aparece en 1966, lo que puede explicar, en parte, el fenómeno, pero Virgilio publicó la mayoría de su obra en los años cuarenta y cincuenta. El caso de Carpentier también resulta revelador. Sus novelas de mayor envergadura, *El reino de este mundo* (1949), *Los pasos perdidos* (1953) y *El siglo de las luces* (1962), fueron publicadas en este periodo, pero su presencia es marginal. En interés de la crítica literaria de esta época ha prevalecido la mirada a la literatura decimonónica y a la figura de Martí.

3.3. La presencia de la literatura cubana en la bibliografía crítica en el período 1968-1989

Durante estos años el IPLN de Cuba aumenta considerablemente y se desmarca de cualquier otra literatura nacional (figuras 4 y 5). De igual forma, hemos realizado un análisis textométrico de las variables título y etiquetas de las referencias de esos años. Como ya hemos comentado, el interés por la literatura del siglo XIX disminuye considerablemente hasta representar un 22 % del total, pero sigue siendo Martí quien acapara la mayoría de las referencias y un factor capital en la presencia de su literatura.

Aumenta el interés por el estudio de los contemporáneos. Tres escritores repuntan con fuerte presencia: Alejo Carpentier, Lezama Lima y Guillermo Cabrera Infante. En el caso de Carpentier, se produce durante esta época otro efecto de autor, pues el 21% de las referencias bibliográficas de este periodo tiene como objeto de estudio principal su obra.

El análisis textométrico nos permite identificar otro fenómeno: el efecto de recuperación de un autor olvidado. Se recuperan escritores que hasta ahora no habían tenido atención crítica. Nicolas Guillén, cuya obra más significativa fue producida varias décadas antes (*Sóngoro cosongo* fue publicada en 1931 y *El son entero* en 1947), es en estos años retomado por la crítica. También es muy significativa la recuperación de Gertrudis de Avellaneda.

En este periodo aparece el término *revolución*, tanto en español como en inglés. En casi un 5% de las referencias aparece explícitamente en el título, frente a una casi inexistencia de esta relación entre revolución y literatura en la etapa anterior. *Literatura y revolución en Cuba* (publicado en 1969, autor Fausto Maso) o *The Short Story of the Cuban Revolution* (publicado en 1971, autor Seymour Menton) son algunos de los ejemplos que muestran como el interés por la Revolución Cubana tiene un impacto directo en la dimensión de la presencia de su literatura. A partir de los años 90, los estudios que relacionan la revolución con la literatura continúan aumentando.

3.4. La presencia de la literatura cubana en la bibliografía crítica a partir del año 1990

El año 90 supone otro punto de inflexión muy significativo en la presencia de la literatura cubana. Es el periodo más complejo de analizar, ya que el número de referencias es mucho mayor, pasamos de unas dos mil quinientas a ocho mil. El análisis textométrico de las variables refleja la aparición de nuevos términos y categorías que explican este aumento. Recorrerlas precisaría de un espacio que excede la naturaleza de este artículo, pero vamos a pensar algunas de las más relevantes.

Aparecen con gran peso estadístico el término *exilio*. Hasta los años 90 la relación entre la crítica literaria y la categorización del exilio es marginal. A partir de los años 90 títulos como *Teatro cubano del exilio* (publicado en 1991, autor Julio Matas), *Hilda Perera: Una voz en el exilio* (publicado en 1991, autora Raquel Romeu), *En torno a la poesía cubana de los años 70 fuera de Cuba* (publicado en 1993, autora Grisel Pujalá), *Call My Son Ishmael: Exiled Paternity and Father/Son* (publicado en 1994, autor Benigno Sanchez-Eppler), *Nacionalidad y exilio en la narrativa cubana contemporánea: Reflexiones sobre la narrativa cubana a partir de 1959* (publicado en 1994, autor Emilio Bejel) o *El exilio florido: José Martí en México* (publicado en 1995, autor José Matesanz), muestran el uso variado y constante del exilio en la justificación y en los criterios de los estudios literarios. Casi quinientas referencias aluden explícitamente al exilio en este periodo frente a las menos de treinta del periodo anterior. A partir de esto años también se multiplican los estudios sobre escritores exiliados y se relaciona, con frecuencia, el exilio con su literatura. Es el caso de Reinaldo Arenas, que muere en el año 90 pero cuya bibliografía crítica se produce en su gran mayoría durante este periodo. Los estudios sobre Severo Sarduy, otro de los escritores del exilio, también aumentan considerablemente en este periodo.

Destacamos la aparición de otro efecto de recuperación de un autor, el caso de Virgilio Piñera, cuya obra más significativa es de los años cuarenta y cincuenta. En los dos periodos anteriores los estudios centrados en su obra son escasos. Pasa, durante estos años, de treinta a más de doscientas referencias.

4. CONCLUSIÓN

Las herramientas digitales nos permiten aplicar métodos cuantitativos a las grandes bases de datos bibliográficos que arrojan nuevas perspectivas sobre el estudio de la presencia de las literaturas nacionales en los trabajos científicos. El acceso a la *MLA International Bibliography*, la mayor base de datos bibliográficos del mundo sobre literaturas en lenguas modernas, y la posibilidad de descargar las referencias en extensiones informáticas convencionales, nos han permitido ampliar las posibilidades de análisis automáticos. La aplicación de una lectura distante y de una perspectiva criticométrica en referencias masivas muestra que la presencia de las literaturas nacionales en la bibliografía crítica está correlacionada y cumple causalidad con factores y variables de distinta naturaleza. Hemos introducido el Índice de Presencia de una Literatura Nacional (IPLN),

definido en relación con la población del país y cuyo cálculo y análisis temporal nos ha demostrado que la distribución de la presencia de las literaturas nacionales es irregular y no responde a una lógica poblacional. Los datos muestran los casos excepcionales de sobrerrepresentación de las literaturas cubana, argentina y chilena y de subrepresentación de las literaturas ecuatoriana y venezolana. Abrimos, por tanto, nuevas posibilidades de trabajo en el campo de las humanidades digitales que tengan como objetivo explicar por qué se producen estas variaciones.

En la última parte de este trabajo nos hemos centrado en el caso más excepcional de entre los nueve países estudiados, el de la literatura cubana. Combinando métodos cuantitativos y textométricos en miles de referencias bibliográficas cuyo objeto de estudio es la literatura cubana, hemos identificado algunas prácticas que explican el desproporcionado interés crítico hacia su literatura, muy superior al de cualquier otra literatura en lengua española del continente. El efecto de autor –proporción desmesurada de referencias bibliográficas centradas en un único escritor– en José Martí, antes de la Revolución y en el primer decenio revolucionario, caracteriza el interés crítico por la literatura cubana. El posicionamiento de Martí como el único clásico literario, una figura ejemplar y representativa de los cambios políticos en la isla acapara el espacio bibliográfico hasta los años setenta, aunque en los periodos posteriores sigue estando también presente. A partir de los años setenta destaca la recuperación de autores que escribieron su obra más significativa bastante antes, especialmente Alejo Carpentier y la aparición de una asociación explícita entre *Revolución* y literatura. En las publicaciones posteriores a 1990 resulta especialmente significativa, además de los procesos anteriores, la relación de los estudios literarios con el exilio. Se evidencia así la causalidad entre los procesos políticos y la presencia de la literatura.

Consideramos que los resultados de este trabajo muestran, por un lado, la pertinencia y el potencial de aplicar la combinación de métodos cuantitativos y textométricos en los corpus a disposición de los investigadores en humanidades y, por el otro, en el caso particular del estudio de la presencia de las literaturas nacionales en la bibliografía crítica, la oportunidad de avanzar en nuevas vías de investigación que nos aporten perspectivas enriquecedoras en la comprensión de este campo, que no podríamos implementar sin el uso de herramientas digitales.

REFERENCIAS BIBLIOGRÁFICAS

- Bhabha, H. K. (1994). *Nation and Narration*. Londres: Routledge.
- Bloom, H. (1995). *El canon occidental*. Barcelona: Anagrama.
- De la Nuez, I. (1998). *La balsa perpetua. Soledad y conexiones de la cultura cubana*. Barcelona: Editorial Casiopea.
- Ferrer, C. (2011). El boom hispanoamericano: del texto a la pantalla. En S. Juan-Navarro (Ed.), *Nuevas aproximaciones al cine hispánico: Migraciones temporales, textuales y étnicas en el bicentenario de las independencias iberoamericanas (1810-2010)* (pp. 79-102). Barcelona: Promociones y Publicaciones Universitarias.

- _____ (2015). Digital Humanities, Big Data, and Literary Studies: Mapping European Literatures in the 21st Century. *Rupkatha Journal: On Interdisciplinary Studies in Humanites*, 11(1). Recuperado de http://rupkatha.com/V7/n1/01_Digital_Humanities_Big_Data.pdf el 02/04/2020.
- _____ (2017a). Les femmes, sont-elles solubles dans le prix Nobel? Les institutions, les lauréates et les oubliées 1901-2016. *M@gm@*, 15(3). Recuperado de http://www.magma.analisisqualitativa.com/1503/article_09.htm el 01/12/2019.
- _____ (2017b). Étudier la transmission littéraire à l'ère du numérique: des grands écrivains à l'analyse des cocitations. *Trans—Revue de littérature générale et comparée*. doi:[10.4000/trans.1719](https://doi.org/10.4000/trans.1719).
- Heiden, S. y Pincemin, B. (2008). Qu'est-ce que la textométrie ? Présentation. *Textométrie*. Recuperado de <http://textometrie.ens-lyon.fr/spip.php?rubrique80> el 10/05/2020.
- Heiden, S., Magué, J. P. y Pincemin, B. (2010). TXM: Une plateforme logicielle open-source pour la textométrie - conception et développement. En *10th International Conference on the Statistical Analysis of Textual Data - JADT 2010* (pp. 1021-1032). Rome: Edizioni Universitarie di Lettere Economia Diritto.
- Moretti, F. (2009). Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740-1850). *Critical Inquiry*, 36(1), 134-158. doi:[10.1086/606125](https://doi.org/10.1086/606125).
- _____ (2013). *Distant Reading*. London, New York: Verso.
- Price, D. (1963). *Little Science, Big Science*. New York: Columbia University Press.
- Ruth-Ellen, St. O. (2013). Compte rendu de Moretti (Franco), *Distant Reading*. *COnTEXTES*. Recuperado de <http://journals.openedition.org/contextes/5870> el 03/05/2020.
- Rojas, R. (2000). *Un banquete canónico*. México: Fondo de Cultura Económica.
- Santí, E. M. (1996). *Por una poliliteratura. Literatura hispanoamericana e imaginación política*. México: UNAM.
- Schöch, Ch. (2017). Topic Modelling Genre: An Exploration of French Classical and Enlightenment Drama. *Digital Humanities Quarterly*, 11(2). Recuperado de <http://digitalhumanities.org/dhq/vol/11/2/000291/000291.html> el 15/02/2020.