doi: 10.15446/rcp.v25n2.50742

Los Secretos de *Cien Años de Soledad*: Una Aproximación Estilométrica para la Investigación en Psicolingüística*

JORGE IVÁN VÉLEZ
Universidad Nacional de Australia, Canberra, Australia
Universidad del Norte, Barranquilla, Colombia

FERNANDO MARMOLEJO-RAMOS
Universidad de Estocolmo, Estocolmo, Suecia

Cómo citar este artículo: Vélez, J. I. & Marmolejo-Ramos, F. (2016). Los secretos de *Cien años de soledad*: una aproximación estilométrica para la investigación en psicolingüística. *Revista Colombiana de Psicología, 25*(2), 265-288. doi: 10.15446/rcp.v25n2.50742

La correspondencia relacionada con este artículo debe dirigirse al Dr. Fernando Marmolejo-Ramos, e-mail: fernando.marmolejo.ramos@psychology.su.se. Gösta Ekman Laboratory, Department of Psychology, Stockholm University, Frescati Hagväg 9A, Stockholm 114 19, Sweden.

ARTÍCULO DE INVESTIGACIÓN CIENTÍFICA
RECIBIDO: 20 DE MAYO DEL 2015 - ACEPTADO: 26 DE FEBRERO DEL 2016

* JIV fue parcialmente financiado por The Eccles Scholarship in Medical Sciences, The Fenner Merit Scholarship y The Australian National University (ANU) High Degree Research Scholarship. Este artículo está dedicado a la memoria del escritor Gabriel García Márquez.

Resumen

De acuerdo con la estilística y la crítica literarias, *Cien Años de Soledad* de Gabriel García Márquez, se caracteriza por aludir constantemente a los personajes de la historia y narrar los eventos en tono neutro. En este artículo se utilizan métodos estilométricos para ratificar dichas afirmaciones y proveer nuevas visiones sobre la novela. Estos métodos incluyen, entre otros, el conteo de palabras y de frases, la construcción de árboles de consenso, el cálculo de la polaridad de las oraciones, e índices para cuantificar la complejidad y el nivel de concreción del texto. Los resultados indican la tendencia del autor a emplear, frecuentemente, palabras abstractas y palabras referentes a objetos con los que se puede interactuar físicamente, para producir el efecto lingüístico propio del realismo mágico. Dada la importancia de los hallazgos, se plantean algunas ideas acerca de las implicaciones que la metodología puede tener en áreas de la psicolingüística y de la psicología cognitiva.

Palabras clave: estilometría, realismo mágico, psiconarratología, psicología cognitiva, cognición corporeizada.

The Secrets of One Hundred Years of Solitude: *A Stylometric Approach for Psycholinguistic Research*

Abstract

According to stylistics and literary criticism, *One Hundred Years of Solitude* by Gabriel Garcia Márquez is characterized by constant allusion to historical characters and a neutral tone in the narration of events. In this article, stylometric methods are used to support these statements and provide new insights into the novel. These methods include, among others, word and phrase count, the construction of consensus trees, the calculation of the polarity of the sentences, and indexes to quantify the complexity and the level of concreteness of the text. Results show the tendency of the author to frequently employ abstract words and words referring to objects with which one can interact physically, to produce the linguistic effect of magical realism. Given the importance of the findings, the article poses some ideas about the implications that the methodology may have for psycholinguistics and cognitive psychology.

Keywords: stylometry, magic realism, psycho-narratology, cognitive psychology, embodied cognition.

Os Segredos de Cem Anos de Solidão: *uma Aproximação Estilométrica para a Pesquisa em Psicolinguística*

Resumo

De acordo com a estilística e a crítica literárias, *Cem anos de solidão*, de Gabriel García Márquez, caracteriza-se por fazer alusão constante aos personagens da história e narrar os eventos sobre o neutro. Neste artigo, utilizam-se métodos estilométricos para validar essas afirmações e oferecer novas visões sobre o romance. Esses métodos incluem, entre outros, a contagem de palavras e de frases, a construção de árvores de consenso, o cálculo da polaridade das orações e os índices para quantificar a complexidade e o nível de concreção do texto. Os resultados indicam a tendência do autor a empregar, com frequência, palavras abstratas e palavras referentes a objetos com os quais se pode interagir fisicamente para produzir o efeito linguístico próprio do realismo mágico. Tendo em vista a importância dos achados, propõem-se algumas ideias acerca das implicações que a metodologia pode ter em áreas da psicolinguística e da psicologia cognitiva.

Palavras-chave: estilometria, realismo mágico, psiconarratologia, psicologia cognitiva, cognição corporizada.

TRADICIONALMENTE, LA apreciación literaria de un texto narrativo se ha hecho a través de aproximaciones cualitativas, en las que el crítico literario presenta una lista de características textuales, determinadas subjetivamente y que pueden o no ser compartidas por otros críticos. Los textos narrativos pueden calificarse, entre otros rasgos, de acuerdo a su género, estilo, tipo de narrador, modo de enunciación, modo temático y tonalidad afectiva. El estudio y la interpretación de las características textuales y lingüísticas de estos se conocen como estilística literaria (Ducrot & Schaeffer, 1972/1998). Por su característica cualitativa, los estudios de estilística literaria pueden dar cabida a algunas discrepancias en la calificación y las clasificaciones otorgadas por los críticos literarios a una obra. Un ejemplo de esto es *El maestro y Margarita*, novela escrita por Mijaíl Bulgákov entre 1928 y 1940, y publicada en 1966. Respecto al narrador usado en esta obra, Gurevich (2003) señala que, mientras algunos críticos sugieren que Iván Bezdomny (uno de los personajes en la obra) narra la historia, basado en lo que le cuenta Woland (el maestro; otro personaje de la historia), otros indican que lo que Iván cuenta es producto de alucinaciones debidas a su enfermedad. La estilometría se presenta, entonces, como una herramienta para analizar el estilo literario, desde un ángulo cuantitativo y ayudar a confirmar o refutar los análisis típicos de forma cualitativa (ver Eder, Rybicki, & Kestemont, 2015).

El objetivo de este artículo es determinar si las evaluaciones cualitativas propuestas por críticos literarios respecto a una novela altamente reconocida, *Cien Años de Soledad* (en adelante CAdS), pueden sustentarse cuantitativamente mediante el uso de métodos estilométricos, como los árboles de consenso y los índices de polaridad, comprensibilidad/legibilidad, y de concreción, usados por el autor. Se implementan también varios análisis novedosos (ver Análisis Suplementarios) que solo son posibles gracias a la estilometría y que pueden representar un gran valor para investigaciones en psicolingüística y psicología cognitiva.

La organización del artículo es la siguiente: inicialmente, se presenta de manera breve la novela CAdS, con algunas de sus características más sobresalientes, a confirmar cuantitativamente más adelante; a continuación, se describe la estilometría y su relación con otras áreas de investigación en el campo de la lingüística y la cognición. Finalmente se plantea la discusión de una perspectiva más general de los alcances de método, en diferentes áreas de la psicología.

CAdS y el Realismo Mágico

Una revisión de la obra completa de Gabriel García Márquez, en especial de CAdS, está fuera del alcance de este artículo[1]. Los lectores interesados deben remitirse, por ejemplo, a los trabajos de Giordano (1968), Gullón (1971) y, particularmente, Rosso (2007). A continuación se presentan algunas características de CAdS, que serán corroboradas, posteriormente, usando métodos estilométricos (ver sección Métodos).

En términos literarios, el realismo mágico (en adelante RM) es un género, en el que lo sobrenatural se presenta como mundano y esto, a su vez, como sobrenatural o extraordinario. En otras palabras, los tópicos fantásticos y mitológicos se contextualizan en escenarios de la vida real. En el RM, un narrador describe los eventos y cuenta lo que acontece, de forma imparcial. Otros representantes del RM incluyen las novelas de Mijaíl Bulgákov (e.g., *El maestro y Margarita*, publicada en 1966) y de Haruki Murakami (e.g., *La caza del carnero salvaje*, publicada en 1982). En el contexto de la literatura latinoamericana, varios autores han sido asociados con el RM. Particularmente Jorge Luis Borges (e.g., cuento corto *La casa de Asterión,* publicado en 1947) y Alejo Carpentier (e.g., cuento corto *El reino de este mundo*, publicado en 1949) son figuras reconocidas de este género. Sin embargo, la obra

1 El presente estudio utilizó una versión electrónica de esa obra, tomada de García Márquez, G. (2003). *Cien años de soledad*. Madrid: Debolsillo.

estimada ejemplo canónico del RM es CAds, escrita por el colombiano Gabriel García Márquez, en 1967, y cuyo reconocimiento le valió el Premio Nobel de Literatura en 1982 (otro ejemplo del RM de García Márquez puede hallarse en su cuento corto *Un señor muy viejo con unas alas enormes*, publicado en 1955).

CAds cuenta los sucesos que acontecen de generación en generación en la familia Buendía, cuyo patriarca, José Arcadio Buendía, es el fundador de Macondo (un pequeño pueblo que solo es real en el contexto de la historia y de localización geográfica imprecisa). La obra se caracteriza por una estructura narrativa típica de la tragedia y el mito (Giordano, 1968) por la constante alusión a los personajes (Gullón, 1971). Los nombres de estos son mencionados repetidamente, lo cual recalca la metáfora de la historia y su circularidad. Este método literario facilita la comprensión de CAds, pues provee un trasfondo histórico y genérico.

Un aspecto interesante de CAds es que el autor usa un tono neutro en toda la novela, para lograr que lo extraordinario se mezcle con lo ordinario; es decir, evita dar tonalidades emocionales a su narrativa con el fin de que el lector no cuestione la mezcla (ver Gullón, 1971). Un ejemplo de la transición entre lo real y lo sobrenatural ocurre de una manera fluida en el siguiente extracto (ver Giordano, 1968):

> La reconoció en el acto, y no había nada pavoroso en la muerte, porque era una mujer vestida de azul, con el cabello largo, de aspecto un poco anticuado, y con un cierto parecido a Pilar Ternera en la época en que las ayudaba en los oficios de cocina.

En este fragmento, llama la atención cómo lo fantasmagórico tiene cualidades de lo real (otro texto con este matiz puede observarse en la novela *Otra vuelta de tuerca*, publicada en 1898 y escrita por Henry James; Gullón, 1971). Gracias al tono neutro usado en la novela, que no da calificaciones emocionales de ninguna índole, el encuentro entre lo real y lo mítico resulta casi cotidiano.

La Estilometría y sus Relaciones con la Psicolingüística (Narrativa) y la Psicología Cognitiva

La estilometría o estilística estadística, en su definición original, aborda la aplicación de métodos estadísticos al material léxico hallado en textos de cualquier índole y de los que se desconoce su autoría y época de producción (ver Holmes, 1998; Gómez, 1999). Varios métodos de pruebas de hipótesis (pruebas de bondad de ajuste y pruebas de independencia/diferencia en diseños mono- y multifactoriales) y para la generación de hipótesis (análisis de *cluster* jerárquico y análisis de componentes principales) pueden utilizarse en análisis de datos lingüísticos (ver Gries, 2015). De igual manera, el método estilométrico, frecuentemente, se apoya en los gráficos estadísticos para mostrar los resultados.

Aunque este tipo de análisis no ha estado exento de críticas (e.g., McCarthy, Lewis, Dufty, & McNamara, 2006), el análisis gráfico de textos, mediante técnicas de estilística computacional, ha permitido diferenciar los géneros literarios que un autor o, incluso, varios autores, puede(n) tener (ver Efron & Thisted, 1976; Thisted & Efron, 1987). Por ejemplo, Binongo y Smith (1999) lograron diferenciar entre los ensayos y obras de Oscar Wilde, utilizando métodos gráficos en análisis de componentes principales, y Craig (1999) usó análisis discriminante, para determinar si ciertas obras de las cuales se desconocía su autoría fueron compuestas por el dramático inglés Thomas Middleton. En ambos casos, los investigadores hicieron conteos de marcadores textuales como artículos indefinidos (e.g., un, unas), preposiciones (e.g., de, para) y pronombres (e.g., eso, esa), agrupados con la ayuda de métodos gráficos, e identificaron asuntos estilísticos en relación con las obras literarias estudiadas.

Los marcadores textuales son los índices esenciales que se usan para distinguir aspectos tanto superficiales como profundos del texto. McCarthy et al. (2006) señalan que las conjunciones son índices de la cohesión de un texto y pueden

llegar a clasificarse como positivas-aditivas (e.g., además), negativas-aditivas (e.g., pero), positivas-temporales (e.g., antes) y negativas-temporales (e.g., hasta que). La cohesión y la coherencia de un texto corresponden a niveles textuales de alto orden y están relacionadas con la asignación de significado a unidades más extensas que las frases (ver Marmolejo-Ramos, Elosúa, Gygax, Madden, & Mosquera, 2009). Esto sustenta la idea de que el análisis de palabras ofrece información valiosa en torno a diferentes niveles textuales: acerca del texto superficial (e.g., sintaxis), del texto base (e.g., gramática) y del modelo de situación (e.g., pragmática, ver Marmolejo-Ramos, 2007b).

El estudio de la comprensión y la producción de textos literarios y narrativos ha sido particularmente fructífero, en trabajos de psicolingüística, en un área llamada psiconarratología, y en el área de la cognición corporeizada en psicología cognitiva. La psiconarratología estudia experimentalmente cómo aspectos literarios afectan la comprensión de un texto (ver Bortolussi & Dixon, 2003; Dixon & Bortolussi, 2001), mientras la psicología cognitiva se encarga de determinar aspectos psicológicos y cerebrales que permiten la comprensión y la producción del lenguaje y textos narrativos (ver Mar, 2004, 2011). A pesar de que, tradicionalmente, los procesos relacionados con la comprensión de textos se han tratado independientemente de los procesos de producción, actualmente se plantea que estos dos procesos tienden a compartir áreas cerebrales (Cevasco & Marmolejo-Ramos, 2013). Así, la investigación en estas áreas ha logrado determinar, por mencionar solo algunos ejemplos, que los pensamientos y acciones del personaje central de una historia son más difíciles de comprender cuando el lector no tiene acceso a marcadores textuales de habla indirecta libre (técnica en la que se mezclan las voces del personaje y del narrador; Kotovych, Dixon, Bortolussi, & Holden, 2011), que la creación de suspenso o emociones facilita la construcción de información perceptual mencionada en la narrativa (Kneepkens & Zwaan, 1995; Komeda & Kusumi, 2006; Molinari, Barreyro, Cevasco, & Van den Broek, 2011), que los lectores monitorean simultáneamente aspectos causales y temporales de la narración (Zwaan, Magliano, & Graesser, 1995; valga anotar que en este estudio se usó un cuento corto de García Márquez), y que los lectores simulan, fácilmente, las acciones descritas en el texto, cuando se usan pronombres que involucran al lector (e.g., *tú* tajas el tomate vs. *él* taja el tomate o vs. *yo* tajo el tomate; Ditman, Brunyé, Mahoney, & Taylor, 2010).

Es evidente el vínculo entre el texto literario, el lector y el escritor (ver Poyatos, 1977). Lingüistas como Morier, Grammont, Spitzer y Vossler (ver Ducrot & Schaeffer, 1972/1998) ya habían reconocido la relación entre la estilística literaria y la psicología del autor. Modelos actuales sobre la comprensión de textos reconocen estos aspectos y resaltan que las características textuales y los procesos cognitivos de alto orden del lector son esenciales en dicha comprensión (Marmolejo-Ramos & Cevasco, 2014). Es decir, no solo la comprensión de los textos, sino del lenguaje en general, requiere del uso de procesos de memoria, inferencias y simulación. Este último, en particular, requiere el uso de áreas cerebrales a cargo de sistemas perceptuales y motores (Cevasco & Marmolejo-Ramos, 2013; Marmolejo-Ramos & Cevasco, 2014; Rapp, Komeda, & Hinze, 2011; Wojciehowski & Gallese, 2011).

Como se mencionó anteriormente, si la comprensión y la producción de textos comparten procesos cognitivos y áreas cerebrales, es posible que el análisis estilomérico de textos escritos permita dilucidar características psicológicas del autor. De hecho, ya existe investigación en este sentido y los resultados sugieren que a través del análisis de marcadores sintácticos en un texto escrito, es posible caracterizar la personalidad de su autor (ver Luyckx & Daelemans, 2008; Pennebaker & King, 1999). Por ejemplo, Noecker, Ryan y Juola (2013) usaron métodos estilométricos en un corpus de 145 ensayos, cada uno de ~1400 palabras, escritos por estudiantes, y lograron clasificarlos en cuatro grupos, cada uno de ellos correspondiente a un tipo de personalidad

(extrovertido/introvertido, intuitivo-perceptivo, pensativo-sentimental y conocedor-perceptivo). Con este trabajo, más preciso que investigaciones similares, los autores demostraron que es posible establecer el tipo de personalidad, a partir del estilo de escritura usado por un autor.

Sin embargo, el uso de la estilometría en el estudio de las características psicolingüísticas del autor no ha sido explorado en igual proporción. Los estudios estilométricos se han enfocado en los aspectos lingüísticos del autor, sin tener en cuenta características psicolingüísticas (i.e., interacción entre aspectos lingüísticos y procesos psicológicos). En el estudio de McCarthy et al. (2006) se utilizaron métodos estilométricos que indicaban diferencias en la longitud de las frases, el uso de palabras con alto y bajo nivel de imaginabilidad y los marcadores de cohesión, entre otros temas, en textos de Kipling, Wodehouse y Dickens. Aunque los resultados mostraron la capacidad lingüística de esos autores, estos no dicen mucho de su psicolingüística (cabe anotar que el conteo de palabras de alto y bajo nivel de imaginabilidad sí va en esa línea). Así, futuros trabajos orientados al estudio de la psicolingüística del autor pueden hallar en los métodos estilométricos una vía para identificar rasgos claves en la psicología del lenguaje. Por ejemplo, a través del conteo de palabras con diferentes niveles de imaginabilidad y concreción podría determinarse cómo usan los autores un lenguaje que facilita la simulación de lo descrito en el texto. Incluso, por medio de las palabras con contenido emocional se podría caracterizar el tono que el autor maneja en su narrativa. Adicionalmente, utilizando índices de comprensibilidad/legibilidad, que combinan rasgos como longitud de frases, de palabras, número de palabras no comunes, entre otros, se podría estimar cuán comprensible puede llegar a ser un texto escrito (Gómez, 1999).

A continuación se presentan algunos detalles relacionados con el análisis estilométrico de CAds, con el fin de cuantificar aspectos lingüísticos de la obra (ver sección CAds y el Realismo Mágico), así como para intentar comprender el estilo característico del RM emblemático de García Márquez. Se usarán árboles de consenso o clasificación y los índices para la determinación de la polaridad, la comprensibilidad/legibilidad y la concreción del texto. Los resultados se usarán como variables dependientes en este estudio. Dado que esta es una investigación formulativa no se proponen hipótesis específicas en cuanto el objetivo es explorar el fenómeno de interés (ver Kothari, 2004). Se espera que, una vez cuantificadas las palabras, con distintos grados de concreción y emocionalidad, se pueda determinar el estilo propio del RM. Por cuanto este tipo de palabras se han usado para el estudio de procesos psicológicos relacionados con la producción y la comprensión del lenguaje, determinar su uso permite hacer inferencias acerca del autor y sus efectos en el lector. Los resultados permitirán sugerir algunas ideas para trabajos futuros en estilometría, psicolingüística y psicología cognitiva.

Métodos

Materiales

En el presente trabajo se usó una versión digital de CAds (Biblioteca Nacional José Martí, 2011). El documento se convirtió inicialmente de formato PDF a formato texto (extensión .txt) y posteriormente se subdividió en capítulos de acuerdo a la nomenclatura que se encuentra en la versión digital. Como resultado se obtuvieron 20 archivos de texto. Cada capítulo de CAds se sometió a una serie de procesos (i.e., transformaciones), para facilitar el uso de herramientas computacionales y los análisis estilométricos que se describirán más adelante. Estas transformaciones, en términos generales, incluyen: (a) el cambio de mayúsculas por minúsculas; (b) remoción de signos de puntuación y de los espacios antes y después de los párrafos, frases y palabras y, (c) la eliminación de las palabras sin sentido[2] (Porter, 2001).

2 Algunos ejemplos incluyen las palabras de, la, que, lo, muy y sobre.

Análisis Estadísticos

Contenido emocional en frases. La minería de opinión, también conocida como *sentiment analysis*, es un área que estudia opiniones, sentimientos, evaluaciones, actitudes y emociones de las personas sobre productos, servicios, empresas y temas específicos (Liu, 2012). Una de las tareas principales en minería de opinión es la clasificación de la polaridad (o contenido emocional) de un texto; es decir, la definición de una connotación negativa, neutra o positiva con relación a este. En algunos casos, el análisis de tales connotaciones puede extenderse a la determinación de estados emocionales.

La polaridad de un texto resulta de la comparación de este con estándares preestablecidos (un *corpus*) y utilizando un clasificador. Para el análisis de CAds, se extrajeron las frases de cada capítulo y sobre estas se determinó la polaridad, utilizando un *corpus* en español al que se tiene acceso a través del paquete *sentiment* (Jurka, 2012) de R, y un clasificador de máxima entropía (Go, Bhayani, & Huang, s. f.). El resultado del clasificador es un puntaje que define la polaridad del texto evaluado (0=negativa, 2=neutral, 4=positiva). Aunque el corpus y la interfaz ofrecida por el paquete *sentiment* fueron desarrollados principalmente para el análisis de mensajes en Twitter, el presente estudio es el primero en extender su aplicación a otro tipo de textos.

Palabras con nivel de concreción e interacción cuerpo-objeto. El nivel de concreción refiere a "la medida en que una palabra hace referencia a un objeto, ser animado, acción o material que puede ser experimentado directamente por los sentidos" (Paivio, Yuille, & Madigan, 1968, citados en Vega & Fernández, 2011, p. 172). La lista de palabras propuesta por Vega y Fernández (2011) fue usada como corpus para examinar el contenido de palabras con alto y bajo nivel de concreción. En ese trabajo los autores reportan los *ratings* dados a 730 palabras por un grupo de 150 estudiantes de habla hispana. Al dividir la lista basada en el *rating* mediano, la mitad de estas fueron clasificadas como palabras de baja concreción (i.e., abstractas), y la otra mitad como palabras de alta concreción (i.e., concretas). Así se determinó la frecuencia de palabras concretas y abstractas presentes en cada capítulo y, posteriormente, el porcentaje en cada categoría. Como ilustración, supongamos que 36 de las 365 palabras concretas y 27 de las 365 palabras abstractas están presentes en el capítulo 1; los porcentajes serán 9.86% (36/365) y 7.39% (27/365), respectivamente. Para el cálculo de este porcentaje, el número de veces que cada palabra está presente no se tuvo en cuenta.

La interacción cuerpo-objeto (*body-object interaction* [BOI] en inglés) determina el nivel de facilidad con el que una persona puede interactuar físicamente con el objeto al que refiere una palabra (e.g., es más fácil interactuar con el objeto referido por la palabra *bicicleta* que con aquel de la palabra *arcoíris*; ver Siakaluk et al., 2008). Una lista de 1618 palabras BOI propuesta por Tillotson, Siakaluk y Pexman (2008) se usó para explorar el nivel de BOI en CAds. Puesto que los *ratings* para estas palabras se construyeron con angloparlantes, las palabras seleccionadas fueron traducidas al español. Al hacer una partición de la lista basada en la mediana del *rating*, 815 palabras fueron clasificadas como palabras de BOI bajo (baja interacción cuerpo-objeto) y 803 como palabras con BOI alto (alta interacción cuerpo-objeto). Finalmente, se determinó el porcentaje de palabras con BOI alto o bajo presentes en cada capítulo de manera similar al procedimiento descrito en el caso de niveles de concreción. Así por ejemplo, si 100 palabras de BOI alto aparecieran en el capítulo 1, entonces el 12.45% (100/803) de las palabras en ese capítulo tendrían un BOI alto.

Comprensibilidad/legibilidad del texto. Un índice de comprensibilidad/legibilidad es la medida del "nivel de dificultad de comprensión del sentido de un texto determinado por ciertos factores lingüístico-estilísticos cuantificables" (Gómez Guinovart, 1999, p. 165; e.g., longitud de palabras,

de oraciones y cantidad de palabras no comunes, entre otras). Aunque existen diversos índices de comprensibilidad/legibilidad (e.g., Gómez, 1999; Michalke, Brown, Mirisola, Brulet, & Hauser, 2014), estos pueden variar en las fórmulas para su cálculo, facilidad de interpretación e idioma en el que fueron validados. En este documento, como una aproximación inicial y, a pesar de que solo ha sido validado para textos en sueco, utilizamos el índice LIX por sus características, fácil interpretación y cálculo, para determinar la complejidad/legibilidad de CAds. Dicho índice puede calcularse como $LIX = A/B + 100\ C/A$, donde A es el número de palabras, B es el número de terminaciones (definidas por punto, punto y coma o primera letra mayúscula) y C es el número de palabras largas (de seis o más letras; ver Hjørland, 2006). Si el índice LIX se encuentra entre 20 y 25, se considera que el texto analizado es *muy fácil* de leer; si está entre 30 y 35, que el texto es *fácil* de leer; *medio* para valores entre 40 y 45; *difícil* si está entre 50 y 55, y *muy difícil* si es mayor a 60 (ver My Byline Media, s. f.)[3].

3 A manera de dato histórico, la referencia original del índice LIX puede hallarse en Björnsson, C. H. (1968). Läsbarhet. Stockholm: Liber.

Aspectos computacionales. Todos los análisis fueron realizados en R a través de desarrollos propios y de algunas de las rutinas ya implementadas en los paquetes *tm* (Feiner & Hornik, 2015), *wordcloud* (Fellows, 2014), *stylo* (Eder et al., 2015), *sentiment* (Jurka, 2012) y *koRpus* (Michalke et al., 2014) del mismo programa. El primer paquete permite aplicar métodos de minería de textos, el segundo se usa para la construcción de nubes de palabras, el tercero para generar árboles de consenso, el cuarto para el análisis de palabras emocionales y el último para estimar el índice LIX.

Resultados

Contenido Emocional

Los resultados del análisis de contenido emocional, presentados en la Figura 1, sugieren que el 85% de las frases (n=4603) tienen una polaridad neutral, el 6.2% (n=337) una polaridad positiva y el 8.8% (n=475) una polaridad negativa, y confirman lo descrito, anteriormente, acerca de un narrador que, usando un tono neutro, permite que lo sobrenatural y lo cotidiano se mezclen de forma natural e inadvertida.

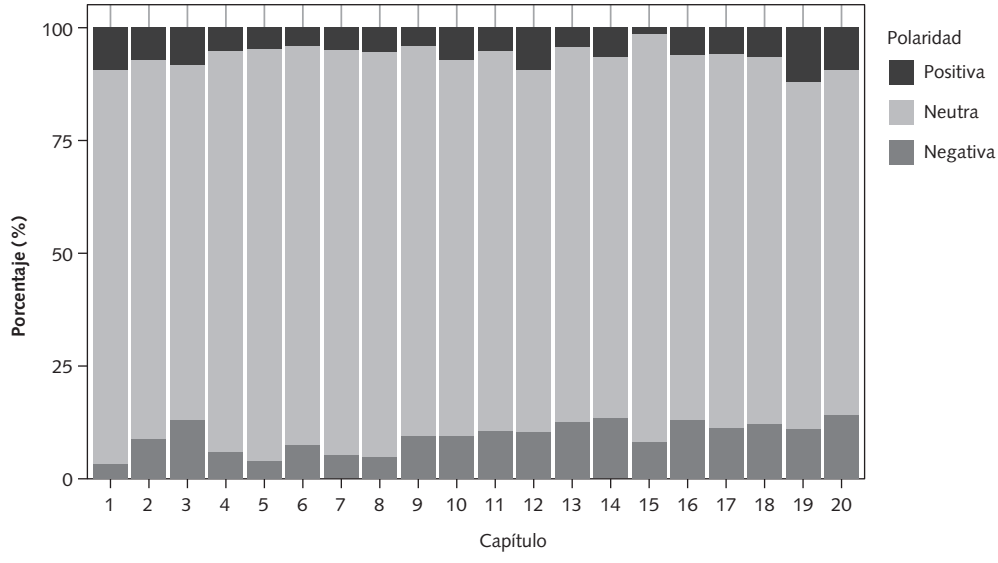

Figura 1. Porcentaje de frases con polaridad en cada uno de los capítulos.

Nivel de Concreción y Palabras de Interacción Cuerpo-Objeto

De acuerdo con la Figura 2a, el autor tiende a usar más palabras de tipo abstracto (o de baja concreción) que palabras de tipo concreto (alta concreción), especialmente en los capítulos 8, 9 y 14 (nótense los picos más altos en la Figura 2b). En los capítulos 1, 5, 6, 7 y 10, por ejemplo, el autor usa más palabras concretas que abstractas. A lo largo de la novela, 12.4% de las palabras son de tipo abstracto y 11.6% son de tipo concreto. Sin embargo, tal diferencia no fue estadísticamente significativa, de acuerdo a una prueba para dos proporciones, z=.11, p=.45. La Figura 2c muestra el porcentaje de palabras de alto y de bajo BOI usadas en cada capítulo. Aunque, en general, la novela parece hacer un uso equivalente de palabras de alto (13.9%) y de bajo BOI (14.2%), parece haber una mayor tendencia hacia el uso de palabras de alto BOI sobre palabras de bajo BOI (ver Figura 2d). Por ejemplo, en los capítulos 4, 12 y 15 hubo una gran diferencia entre el porcentaje de palabras de alto BOI sobre las de bajo BOI. No obstante, la diferencia no fue estadísticamente significativa, de acuerdo a una prueba para dos proporciones, z=- .01, p=.49.

Comprensibilidad del Texto

La Figura 3 sugiere que CAds empieza como un texto *algo difícil* de comprender y termina siendo *muy complejo*. En promedio, CAds se puede considerar un texto difícil de leer (con un valor LIX promedio de 53.03) y se vuelve aún más complejo, a medida que se desarrolla la historia. Por

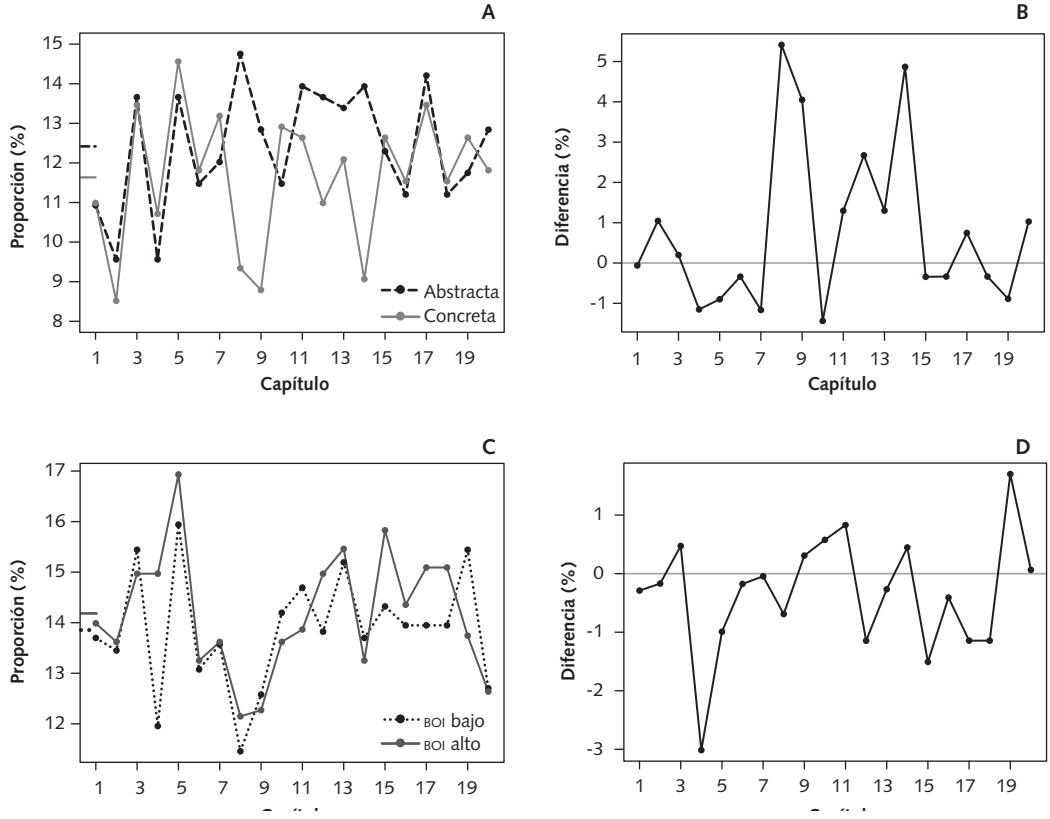

Figura 2. (a) Proporción de palabras con distintos niveles de concreción; (b) valores de diferencia de concreción = abstracta – concreta; (c) efecto BOI a lo largo de los capítulos; y (d) valores de diferencia efecto BOI = BOI bajo – BOI alto.

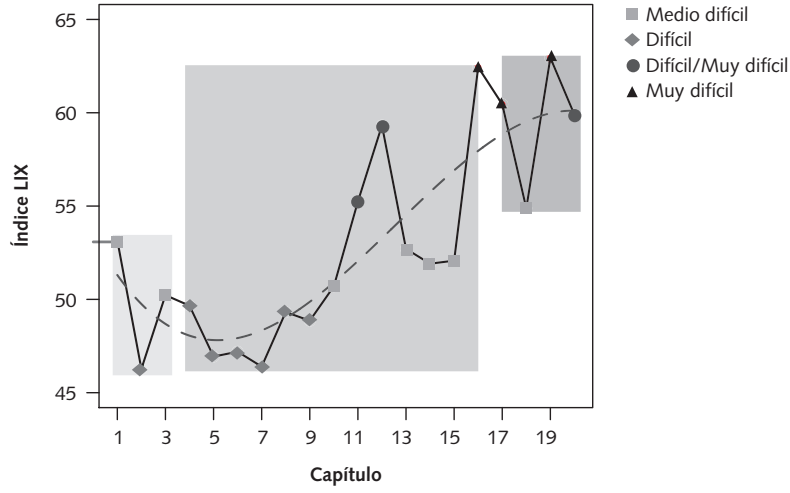

Figura 3. Nivel de comprensibilidad/legibilidad, de acuerdo con el índice LIX para cada uno de los capítulos. Los cuadros grises, de izquierda a derecha, indican los tres grupos temáticos en que se organiza la obra: el establecimiento de Macondo, el desarrollo del pueblo, y su decadencia, respectivamente. La línea puntuada representa la tendencia del índice LIX a lo largo del texto.

ejemplo, mientras los tres primeros capítulos son *medio difíciles* de comprender ($LIX_{promedio}$=49.88), los capítulos que cuentan el desarrollo de Macondo son *difíciles* ($LIX_{promedio}$=51.75), y los cuatro últimos capítulos tienden hacia una complejidad entre *difícil* y *muy difícil* ($LIX_{promedio}$=59.54).

Discusión

Este trabajo tuvo como objetivo demostrar, cuantitativamente, con la ayuda de técnicas de estilometría, afirmaciones cualitativas con respecto a la obra CAds. El método estilométrico permitió determinar que el autor tiende a hacer uso de palabras abstractas y palabras referentes a objetos con los que se puede interactuar físicamente, para producir el efecto lingüístico característico del RM. Aunque también se confirmó que CAds usa una tonalidad emocional neutra (ver Figura 1), los hallazgos más interesantes están relacionados con los niveles de concreción y legibilidad de la obra. Gracias a éstos análisis basados en el conteo de distintos tipos de palabras y/o el análisis de contenido (Holmes, 1998), se halló que en CAds se hace un mayor uso de palabras abstractas que concretas y esto se matiza con el uso de palabras de alto BOI. La correspondencia entre palabras concretas y palabras abstractas es interesante, puesto que corrobora la idea de que los conceptos abstractos se ponen en relación con palabras que se refieren a objetos tangibles con los que es fácil interactuar físicamente.

Los resultados en relación con las palabras BOI son interesantes desde el punto de vista de la relación entre la cognición corporeizada y el lenguaje, pues son indicadores de que las palabras de alto BOI facilitan el procesamiento lingüístico (ver Siakaluk et al., 2008; Xue, Marmolejo-Ramos, & Pei, 2015). Se ha observado que este tipo de palabras requiere la activación de áreas cerebrales encargadas de retener memorias cinestésicas (ver Hargreaves et al., 2012; Siakaluk et al., 2008; Xue et al., 2015). Desde un punto de vista literario, esta parece ser la intención del RM: que el lector entienda lo intangible, a través de analogías entre elementos del mundo real con los que se tiene vasta experiencia. En otras palabras, en el RM se mezcla lo tangible con lo etéreo; lo primero está caracterizado por el uso de palabras de alto BOI y lo segundo por el uso de palabras abstractas. Es importante resaltar que los análisis de concreción y BOI son novedosos en el análisis de corpus de textos literarios, pues hasta ahora se han usado

prioritariamente en investigaciones en psicolingüística experimental, especialmente en el estudio de la cognición corporeizada y del lenguaje. Por lo tanto, y asumiendo que CAds podría considerarse un ejemplo prototípico del RM, los resultados reportados en este trabajo podrían servir como modelos de comparación para estudios literarios de obras categorizadas en este género literario. Hasta la fecha no se había medido cuantitativamente la complejidad de leer CAds. En este sentido, los resultados, gracias al índice LIX (Figura 3), proporcionan una primera aproximación. Como se mostró antes, la complejidad de la obra incrementa a medida que la historia se desenvuelve y en el proceso hay altibajos que caracterizan capítulos específicos. En otras palabras, existe un incremento en la complejidad con la que se expresan los sucesos narrativos (i.e., el establecimiento de Macondo, su desarrollo y, finalmente, su decadencia), que se pone de manifiesto en la construcción de las frases que componen el texto[4].

Ideas para Investigaciones Futuras en el Estudio Estilométrico de Textos Literarios

El acceso a listas de palabras es vital para el análisis estilométrico de textos de cualquiera índole. Un aspecto que no se examinó en este estudio fue el relacionado con el nivel de imaginabilidad de las palabras usadas por el autor. Consideramos que estudios estilométricos orientados a la cuantificación de la imaginabilidad de palabras en textos literarios (McCarthy et al., 2006), pero usando un corpus extenso de palabras en español, constituyen "potenciales líneas de investigación". De manera similar, una futura línea de trabajo consistiría en producir listas extensas de palabras en castellano y construir *ratings* para los niveles de concreción, imaginabilidad, BOI, emocionalidad, y significado, entre otros. Aunque existen ya varias listas en idioma inglés en relación con cada uno de esos aspectos (e.g., Brysbaert, Warriner, & Kuperman, 2014), una lista en español que combine tales *ratings* para cada palabra todavía es una asignatura pendiente. Adicionalmente, la lista de palabras podría evaluarse mediante *ratings* en cuanto a su significancia psicológica y social. Una lista de palabras de este tipo existe en inglés (ver Pennebaker & King, 1999), pero no en español. Dicha lista permitiría análisis más sólidos con respecto a la manera como operan las palabras en los textos, facilitaría la caracterización completa de estos, así como indagar los aspectos psicológicos de los autores.

Los métodos usados en este análisis fueron esencialmente de tipo gráfico, pero el uso de otras técnicas podría ser útil para complementar los análisis. Aunque los análisis gráficos ofrecieron importante información relacionada con los aspectos lingüísticos de interés, los análisis estadísticos formales se pueden usar para corroborar la información gráfica y realizar pruebas de hipótesis. Tal como lo propone Gries (2015), los modelos lineales mixtos y los modelos aditivos generalizados para localización, escala y forma (ver Stasinopoulos & Rigby, 2007) pueden acompañar investigaciones estilométricas. Para el caso de pruebas de hipótesis robustas, métodos como el estadístico tipo ANOVA (ver Noguchi, Gel, Brunner, & Konietschke, 2012) y pruebas basadas en permutaciones (e.g., Marozzi, 2014) serían recomendables.

El tipo de análisis reportado en este documento puede usarse perfectamente para analizar las obras completas de Gabriel García Márquez y así proveer una visión más amplia de su estilo. Es decir, podría crearse una base de datos de versiones digitales de todas sus obras, se podría determinar en qué obras el autor usa más o menos connotaciones emocionales, cuáles obras son más o menos *difíciles* de comprender, e incluso, a través de la técnica conocida como árboles de consenso, se podrían agrupar obras de acuerdo con estas u otras características de interés. Es más, investigaciones con textos literarios podrían usar métodos estilométricos para comparar los estilos característicos de autores clasificados en el género RM (por ejemplo, Gabriel García Márquez vs. Jorge Luis Borges).

4 Un análisis de palabras y frases indicó que en CAds se usan más palabras para construir menos frases, lo cual implica una lectura más compleja a medida que se desarrolla la historia (ver sección Apéndice).

Implicaciones para la Psiconarratología y la Psicología Cognitiva

En relación con las aplicaciones en psicología educativa y cognitiva (psiconarratología), los análisis gráficos reportados en este documento pueden utilizarse no solo en textos narrativos, sino también en textos argumentativos y expositivos. En el primer caso, sería interesante combinar métodos estadísticos, actualmente usados en psiconarratología (ver Bortolussi & Dixon, 2003) con los métodos provistos por la estilometría. En particular, gracias al análisis de contenido sería posible determinar los usos de cierto tipo de palabras, que hacen autores expertos versus autores novatos, para la caracterización del narrador. Igualmente, se podrían generar árboles de consenso y análisis de correspondencia o componentes principales para reunir estos dos grupos de autores, en relación con la complejidad lingüística de la que se hizo uso para la construcción del narrador.

La combinación de métodos estilométricos podría aplicarse a la investigación de textos expositivos y argumentativos escritos por expertos y novatos en un área específica del conocimiento. Recientemente, se ha hallado que estudiantes que leen textos expositivos en un segundo idioma (e.g., cuando el idioma nativo es el castellano y se debe leer un texto en inglés), presentan dificultades al responder preguntas implícitas acerca de aspectos mencionados en el texto y se ha recomendado que, en tales casos, se empiece con preguntas explícitas simples para ir allanando el camino a las preguntas más complejas (Marmolejo-Ramos, Miller, & Habel, 2014). La estilometría de textos expositivos podría ayudar a extraer las palabras de contenido mencionadas más frecuentemente en los textos y así generar, a partir de ellas, conversatorios que ayuden al lector a pasar de un procesamiento superficial del texto a uno más profundo.

En cuanto a los textos de tipo argumentativo, el análisis estilométrico podría usarse en corpus de artículos científicos publicados en distintas áreas científicas (e.g., neurociencias y psicología cognitiva), para identificar el tipo de lenguaje característico de cada área. Tal análisis podría enfocarse también hacia un área de conocimiento, clasificando los artículos de acuerdo al factor de impacto de la revista en la que se han publicado. De tal modo, sería posible investigar si el lenguaje utilizado en los artículos publicados en revistas de alto factor de impacto difiere sustancialmente del usado en los artículos provenientes de revistas de bajo factor de impacto. Los resultados de esos estudios permitirían dar pautas acerca del tipo de lenguaje que los autores deberían usar para incrementar sus posibilidades de publicación en revistas altamente reconocidas.

Recientemente se ha sugerido que la comprensión de acciones, eventos y objetos ficcionales referidos en los textos narrativos, se facilita en la medida en que el lector se apoya en las experiencias sensoriomotoras que ha adquirido en su interacción con el mundo físico real (Rapp et al., 2011; Wojciehowski & Gallese, 2011). Así, es posible que los análisis estilométricos de textos usando palabras que refieran a este tipo de entidades (e.g., palabras concretas, abstractas, emocionales, etc.) puedan crear una representación más fidedigna de los textos y de sus características lingüísticas. En el contexto de la psicología cognitiva y de la educación, tal tipo de análisis podría usarse tanto en los textos narrativos que los alumnos leen, como en sus propias producciones. Por ejemplo, es posible que niños de edad preescolar presten atención a cierto tipo de palabras (e.g., palabras referentes a acciones) en un texto narrativo, para comprender las emociones de los personajes, mientras que niños un poco mayores se enfoquen en otro tipo de palabras (e.g., palabras referentes a estados mentales) al intentar la comprensión. Aunque existen resultados de investigaciones demostrando que niños en edad preescolar comprenden las emociones en textos narrativos acordes a su edad (ver Marmolejo-Ramos & Jiménez, 2006), no hay investigaciones encaminadas a determinar el tipo específico de palabras que facilite la comprensión. Un análisis estilométrico de los textos narrativos que se les proponen a los niños ayudaría a dar ideas

iniciales sobre el tipo de palabras que se podrían correlacionar con una comprensión exitosa de la dimensión emocional.

Investigaciones recientes sobre la escritura de textos narrativos por niños de mayor edad (10 años) han sugerido que, a través de un cuestionario guiado por un adulto experto, es posible ayudar a que los niños produzcan textos más coherentes (ver De Castro & Correa, 2012). Mediante las herramientas estilométricas sería posible cuantificar la mejora de los textos a lo largo de varias sesiones de reescritura. Específicamente, dado que las palabras y las frases proveen información valiosa de la cohesión y la coherencia de un escrito, un análisis estadístico/estilométrico de tales unidades lingüísticas puede proveer nuevas ideas acerca de los procesos cognitivos que operan en el momento de la producción de textos asistida por un adulto.

Es importante mencionar que, dado que los análisis estilométricos dependen en gran medida del corpus de palabras, sería ideal disponer de una lista de palabras usadas en los textos narrativos que leen los niños o que resulte de los textos que escriben. Dicha lista debería poseer *ratings* hechos por niños de distintos grupos de edades, para cada palabra, acerca de su nivel de imaginabilidad, concreción, BOI, emocionalidad, etc. Estos escenarios experimentales, hasta donde se sabe, están aún por investigarse. Hasta el momento se ha asumido, implícitamente, que se trata de niños sin ninguna dificultad de aprendizaje, pero, ciertamente, las ideas propuestas pueden usarse para el estudio de producciones escritas por niños que presentan dificultades de aprendizaje (ver Dolz, Gagnon, Mosquera, & Sánchez, 2013; Marmolejo-Ramos, 2007a).

Conclusiones

Gracias a un análisis estilométrico ha sido posible cuantificar aspectos literarios y estilísticos de la obra *Cien años de soledad*, y proveer nuevos hallazgos en torno a su nivel de legibilidad y al tipo de palabras usadas en la creación del realismo mágico que caracteriza esta novela. La posición adoptada en el presente trabajo es que el método estilométrico puede usarse para complementar investigaciones en áreas como la crítica literaria, la psicolingüística de textos narrativos, expositivos y argumentativos e incluso en psicología cognitiva. Es deseable que las ideas de investigación sugeridas en este documento se materialicen en estudios enfocados en producciones en idioma castellano.

Agradecimientos

Los autores agradecen las sugerencias y comentarios de Santiago Mosquera y Yolima Espinosa sobre versiones previas de este documento.

Referencias

Argamon, S. (2008). Interpreting Burrow's delta: Geometric and probabilistic foundations. *Literary and Linguistic Computing, 23*(2), 131-147. doi: 10.1093/llc/fqn003

Biblioteca Nacional José Martí. (2011). Biblioteca digital. http://goo.gl/tdkePS

Binongo, J. N. G. & Smith, W. A. (1999). A bridge between statistics and literature: The graphs of Oscar Wilde's literary genres. *Journal of Applied Statistics, 26*(7), 781-787. doi: 10.1080/02664769922025

Björnsson, C. H. (1968). *Läsbarhet*. Stockholm: Liber.

Bortolussi, M. & Dixon, P. (2003). *Psychonarratology: Foundations for the empirical study of literary response.* Cambridge: Cambridge University Press.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*(3), 904-911. doi: 10.3758/s13428-013-0403-5

Cevasco, J. & Marmolejo-Ramos, F. (2013). The importance of studying prosody in the comprehension of spontaneous spoken discourse. *Revista Latinoamericana de Psicología, 45*(1), 21-33.

Craig, H. (1999). Authorial attribution and computational stylistics: If you can tell authors apart, have you learned anything about them? *Literary and Linguistic Computing, 14*(1), 103-113. doi: 10.1093/llc/14.1.103

De Castro, D. P. & Correa, M. (2012). Diferentes tipos de incidencias de los procesos cognitivos de revisión sobre la coherencia de textos narrativos: un estudio con niños de 10 años. *Universitas Psychologica, 11*(2), 441-454.

Ditman, T., Brunyé, T. T., Mahoney, C. R., & Taylor, H. (2010). Simulating an enactment effect: Pronouns guide action simulation during narrative comprehension. *Cognition, 115*(1), 172-178. doi: 10.1016/j.cognition.2009.10.014

Dixon, P. & Bortolussi, M. (2001). Prolegomena for a science of psychonarratology. En W. van Peer & S. Chatman (Eds.), *New perspectives on narrative perspective* (pp. 275-287). Albany, NY: S.U.N.Y. Press.

Dolz, J., Gagnon, R., Mosquera, S., & Sánchez, V. (2013). *Producción escrita y dificultades de aprendizaje.* Barcelona: Graó.

Ducrot, O. & Schaeffer, J. M. (1998). *Nuevo diccionario enciclopédico de las ciencias del lenguaje*. Madrid: Arrecife. (Trabajo original publicado en 1972).

Eder, M., Rybicki, J., & Kestemont, M. (2013). Stylo: A package for stylometric analyses. *Computational Stylistics Group.* Recuperado de
https://sites.google.com/site/computationalstylistics/stylo

Eder, M., Rybicki, J., & Kestemont, M. (2015). Stylo: Functions for a variety of stylometric analyses (R package version 0.5.9) [Computer software].

Efron, B. & Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika, 63*(3), 435-447. doi: 10.2307/2335721

Feiner, I. & Hornik, K. (2015). Tm: A framework for text mining applications within R (R package version 0.6) [Computer software].

Fellows, I. (2014). Wordcloud: Pretty word clouds (R package version 2.5) [Computer software].

García Márquez, G. (2003). *Cien años de soledad*. Madrid: Debolsillo.

Giordano, J. (1968). Cien años de soledad. *Revista Iberoamericana, 34*(65), 184-186.

Go, A., Bhayani, R., & Huang, L. (s. f.). *Twitter sentiment classification using distant supervision* (Project Report CS224N). Recuperado de Stanford University website: http://goo.gl/J9xogz

Gómez, J. (1999). Bases lingüísticas y computacionales del procesamiento de la impropiedad estilística y la legibilidad. *Revista Española de Lingüística Aplicada, 1*, 153-173.

Gries, S. (2015). Quantitative methods in linguistics. En J. D. Wright (Ed.), *International encyclopedia of the social and behavioral sciences* (pp. 725-732). Amsterdam: Elsevier.

Gullón, R. (1971). Gabriel García Márquez & the lost art of storytelling. *Diacritics, 1*(1), 27-32.

Gurevich, O. (2003). The master and Margarita: Why can't critics agree on what it means? *Glossos, 4*. Recuperado de http://www.seelrc.org/glossos/index.php

Hargreaves, I. S., Leonard, G. A., Pexman, P. M., Pittman, D. J., Siakaluk, P. D., & Goodyear, B. G. (2012). The neural correlates of the body-object interaction effect in semantic processing. *Frontiers in Human Neuroscience, 6*(22). doi: 10.3389/fnhum.2012.00022

Hjørland, B. (2006). Readability (and legibility). Recuperado de http://goo.gl/Kpobrx

Holmes, D. I. (1998). The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing, 13*(3), 111-117. doi: 10.1093/llc/13.3.111

Jurka, T. (2012). Sentiment. Tools for sentiment analysis (R package version 0.2) [Computer software].

Kneepkens, E. W. E. M., & Zwaan, R. A. (1995). Emotions and literary text comprehension. *Poetics, 23*(1-2), 125-138. doi: 10.1016/0304-422X(94)00021-W

Komeda, H. & Kusumi, T. (2006). The effect of a protagonist's emotional shift on situation model construction. *Memory & Cognition, 34*(7), 1548-1556.

Kothari, C. R. (2004). *Research methodology. Methods and techniques.* New Delhi: New Age International Publishers.

Kotovych, M., Dixon, P., Bortolussi, M., & Holden, M. (2011). Textual determinants of a component of literary identification. *Scientific Study of Literature, 1*(2), 260-291. doi: 10.1075/ssol.1.2.05kot

Le Figaro. (s. f.). Les 100 meilleurs livres de tous les temps. *Le Figaro.* Recuperado de http://goo.gl/fzGKTI

Liu, B. (2012). *Sentiment analysis and opinion mining*. Toronto: Morgan & Claypool.

Luyckx, K. & Daelemans, W. (2008, June). *Using syntactic features to predict author personality from text*. Paper presented at Digital Humanities 2008, Oulu, Finland.

Mar, R. A. (2004). The neuropsychology of narrative: Story comprehension, story production and their interrelation. *Neuropsychologia, 42*(10), 1414–1434.

Mar, R. (2011). The neural basis of social cognition and story comprehension. *Annual Review of Psychology, 62*, 103-134. doi: 10.1146/annurev-psych-120709-145406

Marmolejo-Ramos, F. (2007a). Niños con dificultades en la escuela. El trabajo con textos narrativos como una forma de intervención. *Psychologica, 44*, 97-109.

Marmolejo-Ramos, F. (2007b). Nuevos avances en el estudio científico de la comprensión de textos. *Universitas Psychologica, 6*(2), 331-343.

Marmolejo-Ramos, F. & Cevasco, J. (2014). Text comprehension as a problem solving situation. *Universitas Psychologica, 13*(2), 725-743.

Marmolejo-Ramos, F. & Jiménez, A. T. (2006). Inferencias, modelos de situación y emociones en textos narrativos. El caso de los niños de edad preescolar. *Revista Intercontinental de Psicología y Educación, 8*(2), 93-138.

Marmolejo-Ramos, F., Elosúa de Juan, M. R., Gygax, P., Madden, C., & Mosquera, S. (2009). Reading between the lines: The activation of embodied background knowledge during text comprehension. *Pragmatics & Cognition, 17*(1), 77-107. doi: 10.1075/pc.17.1.03mar

Marmolejo-Ramos, F., Miller, J., & Habel, C. (2014). The role of questions in the comprehension of expository texts. *Higher Education Research & Development, 33*(4), 712-727.

Marozzi, M. (2014). The multisample Cucconi test. *Statistical Methods and Applications, 23*(2), 209-227. doi: 10.1007/s10260-014-0255-x

McCarthy, P. M., Lewis, G. A., Dufty, D. F., & McNamara, D. S. (2006). Analyzing writing styles with Coh-Metrix. *Proceedings of the Florida Artificial Intelligence Research Society International Conference (FLAIRS), 764-770.*

Michalke, M., Brown, E., Mirisola, A., Brulet, A., & Hauser, L. (2014). KORPUS: An R Package for Text Analysis (R package version 0.05-5) [Computer software].

Molinari, C., Barreyro, J. P., Cevasco, J., & Van den Broek, P. W. (2011). Generation of emotional inferences during text comprehension: Behavioral data and implementation through the landscape model. *Escritos de Psicología, 4*(1), 9-17. doi: 10.5231/psy.writ.2011.1803

My Byline Media. (s. f.). Lix Readability Formula: The Lasbarhetsindex Swedish Readability Formula. Recuperado de http://goo.gl/22BQnh

Noecker, J., Ryan, M., & Juola, P. (2013). Psychological profiling through textual analysis. *Literary and Linguistic Computing, 28*(3), 382-387. doi: 10.1093/llc/fqs070

Noguchi, K., Gel, Y. R., Brunner, E., & Konietschke, F. (2012). nparLD: An R package for nonparametric analysis of longitudinal data for factorial designs. *Journal of Statistical Software, 50*(12), 1-23.

Pennebaker, J. W. & King, L. A. (1999). Linguistic style: Language use as an individual difference. *Journal of Personality and Social Psychology, 77*(6), 1296-1312. doi: 10.1037/0022-3514.77.6.1296

Porter, M. (2001). Snowball. Recuperado de http://goo.gl/9OxAQz

Poyatos, F. (1977). Forms and functions of nonverbal communication in the novel: A new perspective of the author-character-reader relationship. *Semiotica, 21*(3/4), 295-338.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Recuperado de http://www.R-project.org/

Rapp, D. N., Komeda, H., & Hinze, S. R. (2011). Vivifications of literary investigation. *Scientific Study of Literature, 1*(1), 122-134. doi: 10.1075/ssol.1.1.13rap

Rosso, C. A. (2007). *Cien años de soledad*. Una ficción que cumple cuarenta años. *El Hombre y la Máquina, 29*, 36-47.

Siakaluk, P. D., Pexman, P. M., Sears, C. R., Wilson, K., Locheed, K., & Owen, W. J. (2008). The benefits of sensorimotor knowledge: Body-object interaction facilitates semantic processing. *Cognitive Science, 32*(3), 591-605. doi: 10.1080/03640210802035399

Stasinopoulos, D. M. & Rigby, R. A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software, 23*(7), 1-46.

Thisted, R. & Efron, B. (1987). Did Shakespeare write a newly discovered poem? *Biometrika, 74*(3), 445-455. doi: 10.2307/2336684

Tillotson, S. M., Siakaluk, P. D., & Pexman, P. M. (2008). Body-object interaction ratings for 1618 monosyllabic nouns. *Behavior Research Methods, 40*(4), 1075-1078. doi: 10.3758/BRM.40.4.1075

Vega, M. & Fernández, A. (2011). Datos normativos de concreción de 730 palabras utilizadas por sujetos de habla castellana. *Psicológica, 33*(2), 171-206.

Wojciehowski, H. C. & Gallese, V. (2011). How stories make us feel. Toward an embodied narratology. *California Italian Studies, 2*(1). Recuperado de http://escholarship.org/uc/item/3jg726c2

Xue, J., Marmolejo-Ramos, F., & Pei, X. (2015). The linguistic context effects on the processing of Body-Object Interaction words: An ERP study on second language learners. *Brain Research, 1613*, 37-48. doi: 10.1016/j.brainres.2015.03.050

Zwaan, R. A., Magliano, J. P., & Graesser, A. C. (1995). Dimensions of situation model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(2), 386-397.

Apéndice*

Análisis Suplementarios

Signos de puntuación, caracteres, palabras, frases y preposiciones. Para cada capítulo se extrajeron los signos de puntuación comúnmente utilizados en español y, posteriormente, se determinó su frecuencia. La selección se realizó usando expresiones regulares en el programa estadístico R (R Core Team, 2014) bajo el criterio de coincidencia completa, y la frecuencia se calculó mediante tablas de contingencia.

En el presente documento, el caracter se considera la unidad lingüística mínima, aunque per se, este no tenga mayor significado. Los caracteres pueden obtenerse a partir de palabras, signos de puntuación y espacios, o combinaciones de estos. Sin embargo, algunas veces solo se considera el primer componente. Así, "reciente, ", "nombre," y "dedo." tienen diez, siete y cinco caracteres, respectivamente, cuando se consideran signos de puntuación y espacios. Este número se reduce a ocho, seis y cuatro cuando solo se consideran las letras que componen cada palabra.

Las palabras se extrajeron luego de convertir las letras mayúsculas a minúsculas y eliminar los signos de puntuación en cada capítulo. Bajo esta convención, las palabras *El* y *nombre,* se transformaron en *el* y *nombre*, respectivamente. Posteriormente, se determinó la frecuencia de cada palabra, utilizando tablas de contingencia, una vez excluidas las palabras sin sentido.

Se definió una frase como el texto que se encuentra a la izquierda de un punto (.) cuando se trata del comienzo de un párrafo, o entre punto (.) y punto (.) en otro caso. Después de convertir las palabras mayúsculas en minúsculas y bajo esta definición, se extrajeron las frases de cada capítulo. A partir de las frases, se determinó la cantidad de estas, por capítulo, el número promedio de palabras por frase y la frase más larga (o más corta) por capítulo y en la novela completa.

Para el caso de las preposiciones[5] se utilizaron 23 palabras propuestas por la Real Academia Española (RAE) y la Asociación de Academias de la Lengua Española (ASALE). La frecuencia de aparición de cada preposición se determinó con la aplicación de un proceso similar al utilizado en el caso de los signos de puntuación.

Análisis de *n*-gramas. Los *n*-gramas son secuencias de *n* elementos continuos en un texto o discurso. Para ilustración, consideremos la siguiente frase: *Por fin, un martes de diciembre,*

* Este es el material suplementario del artículo "Los secretos de *Cien años de soledad*: una aproximación estilométrica para la investigación en psicolingüística" escrito por Jorge I. Vélez y Fernando Marmolejo-Ramos (la bibliografía citada aparece en el artículo).

5 La lista actual incluye las palabras *a, ante, bajo, cabe, con, contra, de, desde, durante, en, entre, hacia, hasta, mediante, para, por, según, sin, so, sobre, tras, versus y vía.*

a la hora del almuerzo, soltó de un golpe toda la carga de su tormento.

Luego de procesar el texto[6], se obtienen las palabras *fin*, *martes*, *diciembre*, *hora*, *almuerzo*, *soltó*, *golpe*, *toda*, *carga*, y *tormento.* Cuando n=1, los unigramas resultantes corresponden a las palabras mencionadas anteriormente. Para n=2, se obtienen nueve pares de palabras consecutivas: *fin martes*, *martes diciembre*, *diciembre hora*, *hora almuerzo*, *almuerzo soltó*, *soltó golpe*, *golpe toda*, *toda carga* y *carga tormento.* Finalmente, para n=3, se obtienen ocho trigramas: *fin martes diciembre*, *martes diciembre hora*, *diciembre hora almuerzo*, *hora almuerzo soltó*, *almuerzo soltó golpe*, *soltó golpe toda*, *golpe toda carga*, y *toda carga tormento*. El proceso es similar para otros valores de n. Una vez se tienen los n-gramas para el texto, el siguiente paso es determinar su frecuencia. Se construyeron n-gramas para n=1, n=2 y n=3 una vez procesado el texto.

Similitud entre capítulos. La novela CAds ha sido ganadora de innumerables premios internacionales y es considerada uno de los mejores 100 libros de todos los tiempos (Le Figaro, s. f.). La novela consta de 20 capítulos no titulados que, de acuerdo con la historia relatada en ellos, podrían organizarse en tres secciones: (a) el establecimiento de Macondo (capítulos 1-3), (b) el desarrollo del pueblo (capítulos 4-16) y (c) su decadencia (capítulos 17-20). A partir de un análisis estilométrico, utilizando un método de clasificación no supervisado, es posible agrupar los capítulos para verificar tal organización. No obstante, también pueden surgir otras organizaciones que podrían explorarse en investigaciones futuras.

Para el análisis de similitud entre capítulos se utilizaron las herramientas implementadas en el paquete *Stylo* (Eder et al., 2013, 2015) de R, a partir de los textos de cada capítulo. Esencialmente, se trata de estimar el grado de similitud entre grupos lingüísticos (i.e., los capítulos) de interés, dadas las semejanzas entre la cantidad de palabras que aparecen en cierto porcentaje de tales grupos y la cantidad de palabras más frecuentes (Eder et al., 2013). El resultado final del análisis se representa gráficamente como un dendrograma (también conocido como *árbol de consenso*). Una de las ventajas de este método, sobre todo de la representación gráfica, es que la interpretación de los resultados es directa y permite la fácil determinación de las particularidades del texto analizado (Eder et al., 2013).

Para los análisis de similitud entre los capítulos de CAds se generaron B=10000 muestras aleatorias *bootstrap* para determinar el árbol de consenso definitivo, se determinaron las frecuencias relativas de palabras, se usó una función de distancia clásica (correspondiente a la distancia de Manhattan aplicada a frecuencias de palabras normalizadas; ver Argamon, 2008) y un valor de asociación de .5. Este último parámetro indica que si la asociación (o similitud) entre capítulos existe, esta debe aparecer en al menos el 50% de los grupos hallados en cada muestra aleatoria *bootstrap*.

Resultados Suplementarios

Caracteres, palabras, frases y preposiciones. Las Tablas A1, A2 y A3 presentan las frecuencias en relación a signos de puntuación, caracteres, palabras, frases y preposiciones. La versión analizada de CAds contiene 8852 comas, 32 punto y coma, 5413 puntos, 1333 pares de guiones, 60 pares de signos de admiración, 35 pares de signos de pregunta y 165 signos de dos puntos. Es importante mencionar que García Márquez no usa las preposiciones *versus* y *vía* puesto que estas tienden a ser más usadas en textos académicos y argumentativos que en textos literarios, mientras que *cabe* y *so* tienden a ser algo arcaicas y quizás más típicas de textos poéticos[7] (ver Tabla A3).

6 Esto es, convertir las mayúsculas en minúsculas y remover los signos de puntuación y las palabras sin sentido.

7 Esta opinión es basada en nuestra experiencia como lectores de textos académicos y no académicos. Sin embargo, tal

Tabla A1
Frecuencia de signos de puntuación por capítulo

	Signo de puntuación								Signo de puntuación							
Capítulo	**,**	**;**	**.**	**-**	**!**	**?**	**:**		**Capítulo**	**,**	**;**	**.**	**-**	**!**	**?**	**:**
1	401	4	223	46	1	1	10		11	423	1	246	69	2	1	4
2	392	1	275	56	1	2	16		12	477	2	221	64	2	2	2
3	439	1	303	51	1	4	17		13	511	1	278	56	8	1	2
4	441	1	301	57	3	2	16		14	471	1	316	57	1	1	3
5	404	3	371	89	1	1	19		15	443	1	295	66	3	2	6
6	365	1	298	80	7	1	6		16	464	1	178	40	1	1	1
7	403	1	334	114	3	3	19		17	458	2	222	56	4	1	4
8	403	1	293	108	3	5	9		18	501	6	240	24	1	1	4
9	427	1	321	142	3	1	13		19	536	1	199	34	8	1	6
10	479	1	306	83	3	3	2		20	414	1	193	41	4	1	6

Nota: En el caso del español los signos ? y ! corresponden a los pares ¿? y ¡! respectivamente.

Tabla A2
Número total de palabras, y palabras únicas por capítulo

Capítulo	Todas	Únicas	Porcentaje (%)	Capítulo	Todas	Únicas	Porcentaje (%)
1	5874	2008	34.2	11	6693	2168	32.4
2	6103	1883	30.9	12	6965	2296	33.0
3	7525	2291	30.4	13	7867	2196	27.9

afirmación debe ser validada a través de un estudio estilométrico.

4	6810	2210	32.5	14	8051	2223	27.6
5	7778	2339	30.1	15	7182	2134	29.7
6	5884	1954	33.2	16	6406	2035	31.8
7	6658	2012	30.2	17	7441	2340	31.4
8	6365	2000	31.4	18	7164	2135	29.8
9	6476	2013	31.1	19	7297	2379	32.6
10	7039	2138	30.4	20	6227	2072	33.3

Nota: El porcentaje de palabras únicas por capítulo se calculó a partir del número de palabras únicas sobre el total de palabras utilizadas.

Tabla A3
Frecuencia de preposiciones

Preposición	Frecuencia	Preposición	Frecuencia
a	3165	hasta	347
ante	42	mediante	18
bajo	48	para	1020
cabe	0	por	1469
con	1985	según	29
contra	86	sin	496
de	8687	so	0
desde	246	sobre	97
durante	87	tras	7
en	3889	versus	0
entre	118	vía	0
hacia	58		

Los resultados que se presentan a continuación en las figuras, hablan por sí solos y este es precisamente el propósito de los métodos estadísticos no supervisados usados en estilometría: que el investigador tenga la oportunidad de explorar los resultados gráficos para buscar peculiaridades o asuntos inesperados en el corpus analizado (Eder et al., 2013). Por ejemplo, las Figuras A1 y A2 presentan algunos conteos de interés. El capítulo con el menor número de caracteres es el capítulo 6 (~28000 caracteres), mientras que el que contiene un mayor número es el capítulo 14 (~37000; Figura A1a); el menor y mayor porcentaje de palabras únicas se obtiene en los capítulos 14 (27.6%) y 1 (34.2%), respectivamente (Tabla A1 y Figura A1b); los capítulos 6 (~20 palabras/frase) y 19 (~36 palabras/frase) presentan el menor y mayor número promedio de palabras por frase (Figura A1c); y el menor y mayor número de frases se presentan en los capítulos 16 (~180 frases) y 6 (~370 frases), respectivamente (Figura A1d).

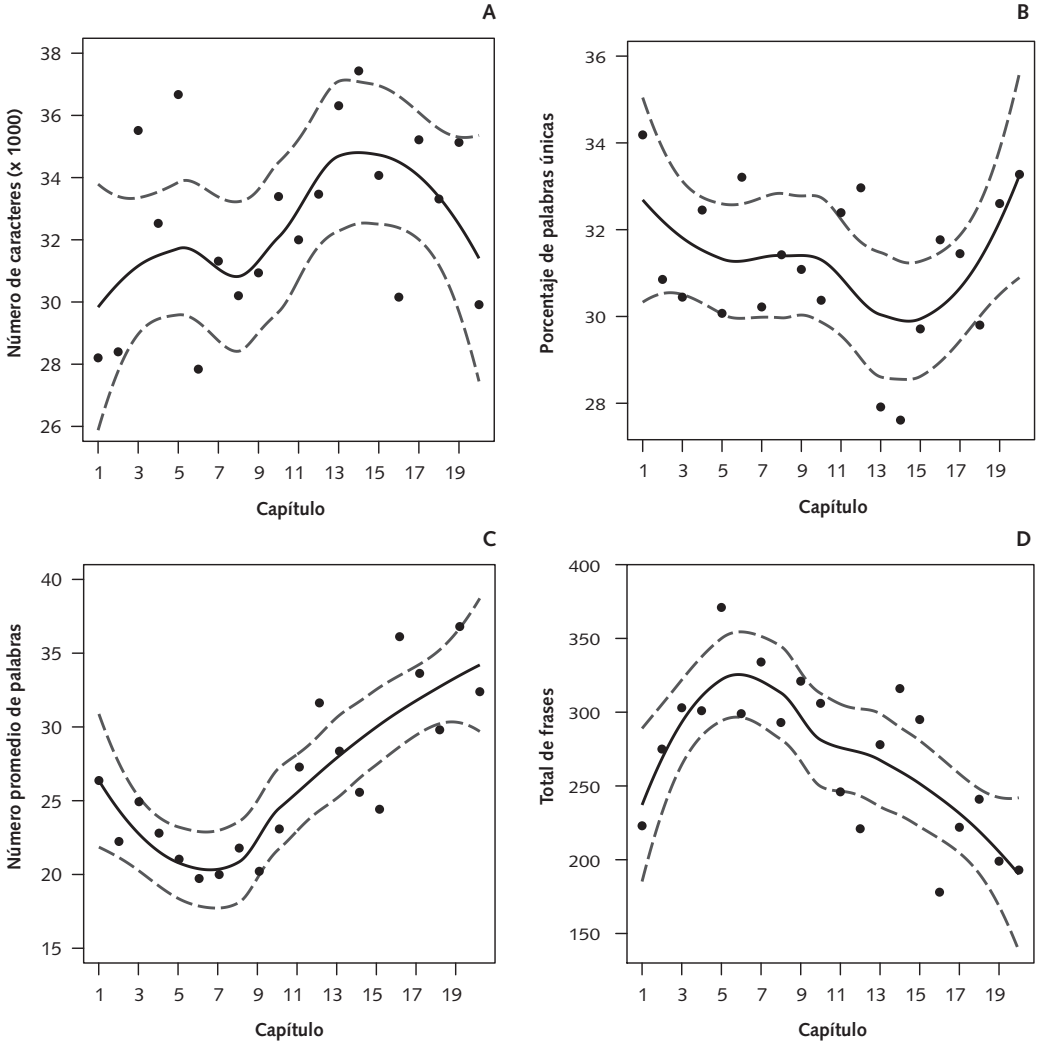

Figura A1. (a) Frecuencia de caracteres, (b) porcentaje de palabras únicas (ver también Tabla A3), (c) número promedio de palabras por frase y (d) total de frases por capítulo. La línea continua corresponde al valor predicho al utilizar un modelo de regresión no paramétrica *loess*; las líneas punteadas corresponden al intervalo de confianza del 95% para dicho valor.

Por otro lado, el comportamiento observado en la Figura A2a sugiere que el número total de frases por capítulo es inversamente proporcional al número promedio de palabras por capítulo. Es decir, Gabriel García Márquez usa un promedio mayor de palabras para elaborar menos frases y viceversa. Los capítulos 1, 2, 5 y 14 están por fuera del intervalo de confianza del 95% (línea roja punteada, Figura A2a) indicando que, en cierta medida, existen algunas diferencias entre estos capítulos y el resto de capítulos en la novela.

Un total de 78 palabras aparecen simultáneamente por lo menos *una vez* en *todos* los capítulos de la novela (Figura A2b). La palabra *Aureliano* es la que se menciona con mayor frecuencia (795 veces), seguida de *Úrsula* (512), *Arcadio* (480), *Casa* (463), *José* (438) y *Buendía* (411). Las palabras comunes de menor frecuencia son *fin* (49), *quedó* (51), *mejor* (53) y *último* (55). El hecho de que la palabra *Aureliano* sea la más frecuente se confirma por los *n*-gramas (ver Figura A4 para *n*=1). La Figura A3 sugiere que los capítulos 1, 6 y 12 presentan el número de caracteres más bajo, mientras los capítulos 5, 13 y 14, los más altos. Igualmente, se observa que la distribución del número de caracteres por frase tiene sesgo positivo en todos los capítulos, y que en el capítulo 16 aparece el mayor número (995 caracteres). Específicamente, se trata de la frase 102 que contiene la célebre "cantaleta" de Fernanda del Carpio a Aureliano y la cual evidencia la magistral narrativa de Gabriel García Márquez cargada de humor[8].

8 Ver http://goo.gl/mKc1Rq

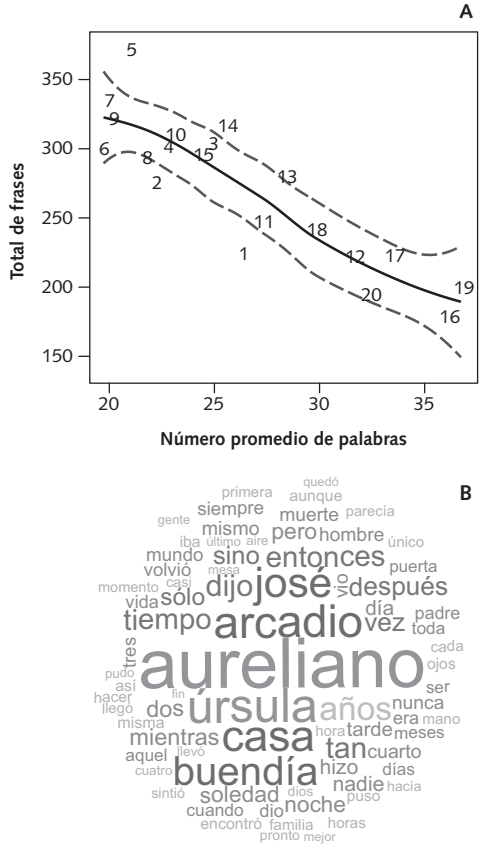

Figura A2. (a) Relación entre el número total de frases y el promedio de palabras por frase, a lo largo de los capítulos (estos, representados por su respectivo número) y (b) nube de palabras representando las palabras con mayor o menor frecuencia. En (b), las palabras en el centro del gráfico aparecen con mayor frecuencia que las de la periferia. Convenciones como en la Figura A1.

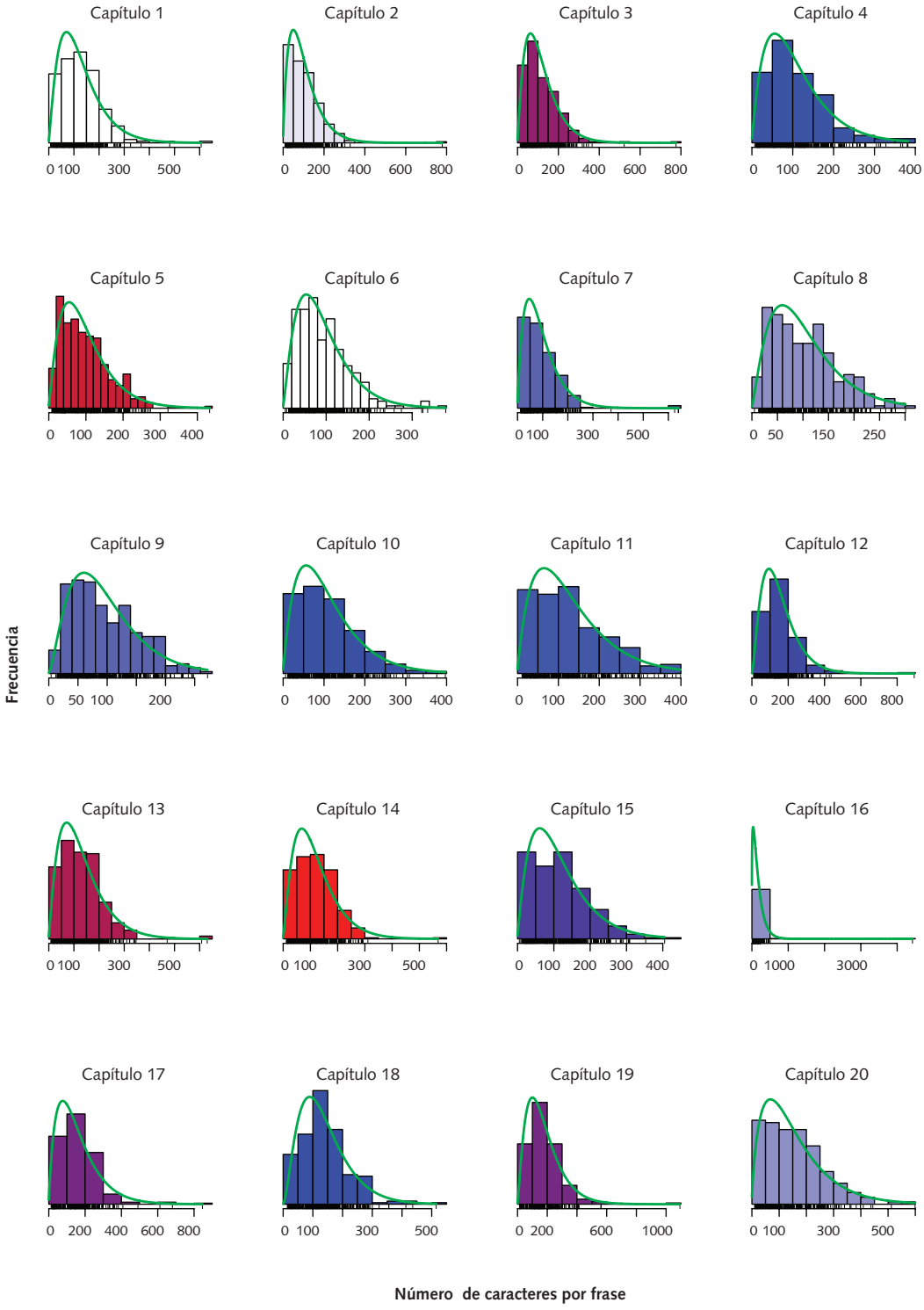

Figura A3. Distribución del número de caracteres por frase para cada capítulo. La escala de colores va desde blanco (valores mínimos), pasa por azul (valores intermedios) y finaliza en rojo (valores más altos) teniendo en cuenta el número total de caracteres por capítulo. La línea de color verde representa la función de distribución de probabilidad de una distribución Binomial Negativa ajustada a los datos.

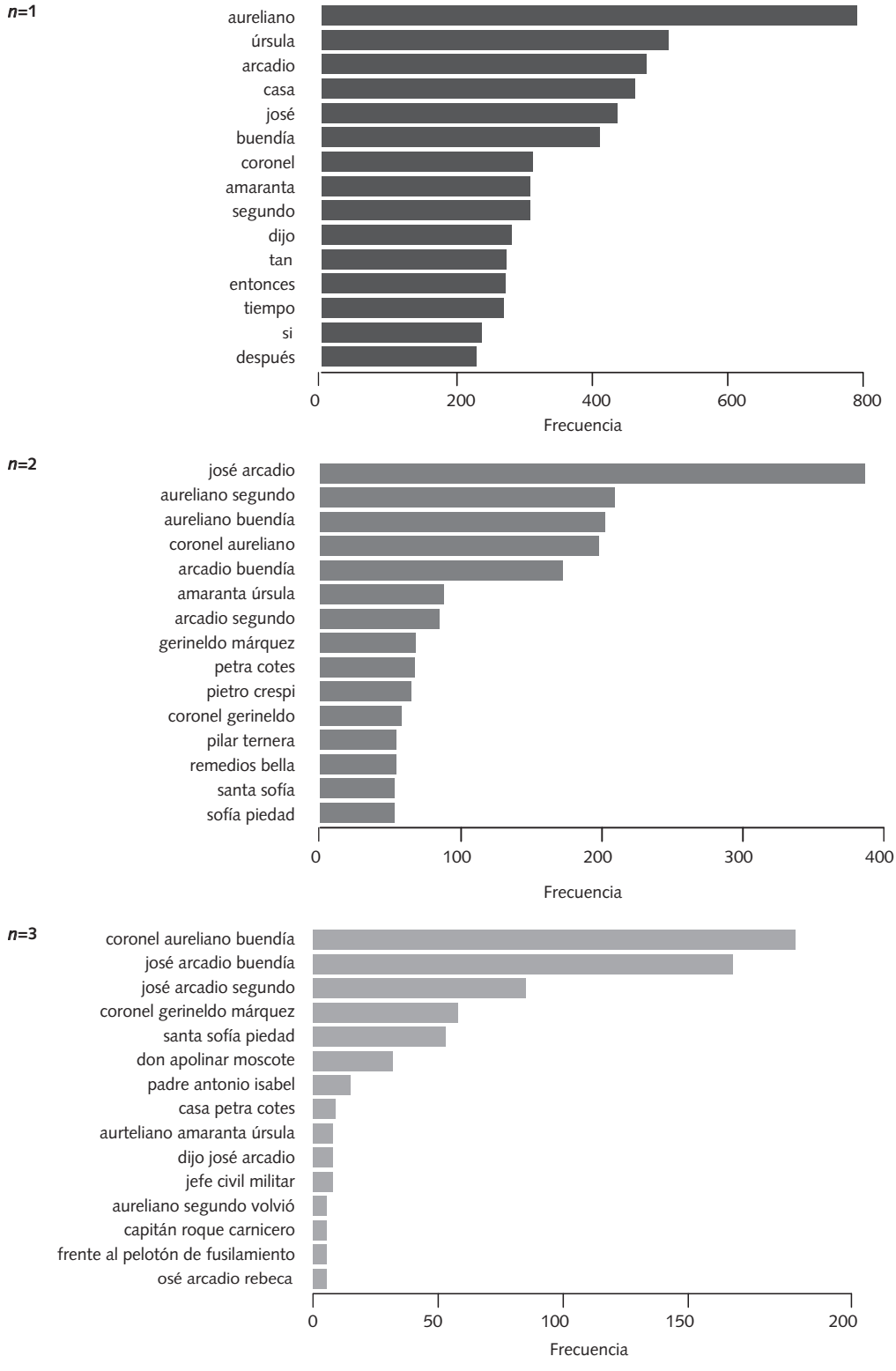

Figura A4. Los 15 *n*-gramas más repetidos y su frecuencia de aparición a lo largo del libro. Los casos de *n*=2 y *n*=3 se incluyen para efectos de comparación.

Similitud entre capítulos. En la Figura A5 se presenta el árbol de consenso obtenido una vez realizado el análisis a partir de palabras. Este resultado es fascinante y despierta un gran interés pues muestra la continuidad temática que caracteriza a de *Cien años de soledad*. El árbol de consenso apoya la idea de que los capítulos 1 a 3 giran alrededor del establecimiento de Macondo, los capítulos 4 a 16 tratan del desarrollo del pueblo, y los últimos capítulos narran su decadencia. Puede verse que, aun cuando el capítulo 3 se aleja de los capítulos 1 y 2, los dos últimos aparecen cerca. Adicionalmente, a pesar de que el capítulo 4 parece *independiente* de los relacionados con el desarrollo del pueblo, los capítulos 5-9, 10-11 y 12-14 aparecen agrupados; incluso los capítulos 15 y 16, aunque están en ramas distintas, se acerca el uno al otro. Finalmente, si bien el capítulo 17 se encuentra desprendido de los otros tres capítulos restantes, sí se halla cerca de estos últimos.

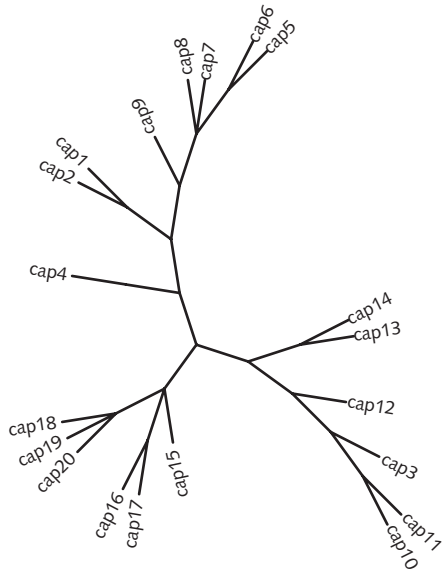

Figura A5. Árbol de consenso resultante del análisis estilométrico por palabras, para los capítulos de CADS.