ONOMÁZEIN
Revista de lingüística, filología y traducción

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
FACULTAD DE LETRAS

Identificación de cambios en el estilo de escritura literaria con aprendizaje automático

Identification of changes in literary writing style using machine learning

Germán Ríos-Toledo
Tecnológico Nacional de México | Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET)
México

Noé Alejandro Castro-Sánchez
Tecnológico Nacional de México | Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET)
México

Grigori Sidorov
Instituto Politécnico Nacional (IPN)
México

Juan-Pablo Posadas-Durán
Instituto Politécnico Nacional (IPN)
México

ONOMÁZEIN 46 (diciembre de 2019): 102-128
DOI: 10.7764/onomazein.46.04
ISSN: 0718-5758

Germán Ríos-Toledo: Tecnológico Nacional de México/Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), México. | E-mail: german_rios@hotmail.com
Noé Alejandro Castro-Sánchez: Tecnológico Nacional de México/Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), México.
Grigori Sidorov: Centro de Investigación en Computación (CIC), Instituto Politécnico Nacional (IPN), México.
Juan-Pablo Posadas-Durán: Escuela Superior de Ingeniería Mecánica y Eléctrica, Unidad Zacatenco (ESIME Zacatenco), Instituto Politécnico Nacional (IPN), México.

Fecha de recepción: marzo de 2018
Fecha de aceptación: julio de 2018

Resumen

Esta investigación tiene como objetivo identificar cambios en el estilo de escritura a través del tiempo de 7 autores de novelas de habla inglesa. Para cada autor se realizó una organización de las novelas de acuerdo a la fecha de publicación. Las novelas se clasificaron en tres etapas denominadas *inicial*, *intermedia* y *final*; cada etapa contiene 3 novelas. Entre dos etapas consecutivas existe por lo menos 2 años de separación entre las fechas de publicación de las novelas. Para resolver el problema de detección de cambios en el estilo de escritura a través del tiempo se propone utilizar un enfoque basado en aprendizaje automático supervisado. Se crearon modelos de espacio vectorial a partir de las frecuencias de uso de n-gramas de distintos tipos y longitudes. Además, se utilizó el algoritmo de Análisis de Componentes Principales (Principal Component Analysis, PCA) como método de selección de n-gramas. La solución se abordó como un problema de clasificación utilizando los algoritmos de Máquinas de Soporte Vectorial (Support Vector Machine, SVM), Naive Bayes Multinomial (Multinomial Naive Bayes, MNB), Regresión Logística (Logistic Regression, LG) y Liblinear como clasificadores. La métrica para medir la eficiencia de los algoritmos de aprendizaje fue la exactitud (*accuracy*). La investigación mostró cambios significativos en cinco de los autores con una exactitud promedio de entre 70% y 80% en los distintos tipos de n-gramas.

Palabras clave: detección de cambios de estilo a través del tiempo; n-gramas; n-gramas sintácticos; modelo espacio vectorial; cambio de estilo; aprendizaje automático.

Abstract

This research aims to identify changes in the writing style over time of 7 authors of English-speaking novels. For each author, an organization of the novels was carried out according to the date of publication. The novels were classified in three stages called initial, intermediate and final; each stage contains 3 novels. Between two consecutive stages there are at least 2 years of separation between the publication dates of the novels. To solve the problem of detecting changes in writing style over time, it is proposed to use a supervised automatic learning-based approach. Vector space models were created from the frequencies of use of n-grams of different types and lengths. In addition, the algorithm of Principal Component Analysis (PCA) was used as the n-gram selection method. The solution was addressed as a classification problem using the Vector Support Machine algorithms (Support Vector Machine, SVM), Naive Bayes Multinomial (Multinomial Naive Bayes, MNB), Logistic Regression (LG)

and Liblinear as classifiers. The metric to measure the efficiency of the learning algorithms was accuracy. The research showed significant changes in five of the authors with an average accuracy between 70% and 80% in the different types of n-grams.

Keywords: detection of style changes over time; n-grams; syntactic n-grams; vector space model; style change; machine learning.

1. Introducción

En el campo del análisis de textos, el estilo se refiere a aquellos elementos lingüísticos que independientemente del contenido del documento persisten a lo largo de todos los trabajos de un autor (Uzuner y Katz, 2005). Para realizar un análisis de estilo de escritura, Gamon y Grey (2004) sugieren el uso de elementos del lenguaje independientes del tópico del documento. Este tipo de análisis se centra principalmente en la forma del texto más que en su contenido, por lo que es recomendable utilizar elementos lingüísticos estructurales como las categorías gramaticales de las palabras, palabras funcionales, longitud de oraciones y la información sintáctica presente en el texto. Turell y Gavaldà (2013) indican que los sociolingüistas han demostrado durante décadas que los lenguajes están en constante cambio y que son intrínsecamente variables en todos sus niveles; en otras palabras, la producción lingüística de un solo hablante o escritor generalmente mostrará alguna variación.

Determinar estos cambios resulta importante en áreas como el diagnóstico de enfermedades neurológicas como la demencia (Hirst y Feng, 2012), así como también en la atribución de textos en disputa (Sidorov y otros, 2014; Björklund, 2016), creación de perfiles de autor (Argamon y otros, 2009; Posadas-Durán y otros, 2015) e identificación de autores (Luyckx y Daelemans, 2008a; Brocardo y otros, 2013). Otra de las posibles aplicaciones de la detección de cambio estilo es la predicción de la personalidad del autor (Luyckx y Daelemans, 2008b), la detección de tendencias a la depresión (Rude y otros, 2004) y en el aprendizaje de una segunda lengua (Yoon y Bhat, 2012).

Para realizar un análisis de estilo es necesario establecer cuáles serán los elementos lingüísticos a utilizar, como, por ejemplo, palabras frecuentes, los signos de puntuación, la longitud de oraciones, la complejidad sintáctica, etc. En esta investigación se propone el uso de elemento conocido como n-grama. Un n-grama es una secuencia de **n** elementos; en el caso de textos, estos elementos pueden ser caracteres, palabras, etiquetas gramaticales y sintácticas.

Los n-gramas de caracter son independientes del lenguaje; pueden capturar elementos como prefijos, sufijos, signos de puntuación, contracciones, letras capitales, etc. Han sido utilizados con muy buenos resultados en tareas como atribución de autoría (Peng y otros, 2003; Kešelj y otros, 2003), detección de plagio (Barrón-Cedeño y Rosso, 2009), detección de tópico (Peng y otros, 2004), entre otros. Los n-gramas de palabras se han utilizado en detección de plagio (Barrón-Cedeño y Rosso, 2009), en clasificación de textos basado en tópico (Peng y otros, 2004), en atribución de autoría (Sanderson y Guenter, 2006; Coyotl-Morales y otros, 2006).

Los n-gramas de etiquetas gramaticales POS (*part of speech*) se forman con las categorías gramaticales de las palabras: verbos, sustantivos, pronombres, adjetivos, adverbios, etc. Se han utilizado en atribución de autoría (Kukushkina y otros, 2001; Koppel y Schler, 2001; Diederich y otros, 2003; Luyckx y Daelemans, 2005; Zhao y otros, 2007), en reconocimiento de autores (Uzuner y Katz, 2005; Hirst y Feiguina, 2007). La información puramente sintáctica

generalmente ha dado resultados inferiores a los enfoques léxicos tradicionales en términos de eficiencia pura de atribución (Fuller y otros, 2013). Pero existen trabajos que indican que combinar la información sintáctica con otro tipo de información produce mejores resultados: en atribución de autoría (Baayen y otros, 1996; Kaster y otros, 2005; Raghavan y otros, 2010), en verificación de autoría (Halteren, 2007).

Un problema de clasificación consiste en asignar una categoría a un elemento partiendo de un conjunto de categorías definidas previamente, analizando la similitud entre el elemento a clasificar y ejemplos de elementos de las categorías. El problema de clasificación se puede resolver mediante algoritmo de aprendizaje supervisado, como, por ejemplo, árboles de decisión, regresión logística (LG) y máquinas de soporte vectorial (SVM). Para medir el desempeño de estos algoritmos existen métricas como precisión, sensibilidad y exactitud (*precision*, *recall* y *accuracy* respectivamente).

La investigación mostró cambios significativos en cinco de los autores con una exactitud promedio de entre 70% y 80% en los distintos tipos de n-gramas. Sin embargo, la informacion sintáctica de dos autores no permitió identificar variaciones importantes entre las etapas, por lo que en estos casos la tasa de exactitud no superó el 50% en la mayoría de los experimentos.

El resto del artículo está organizado de la siguiente manera: la sección 2 describe trabajos en el estado del arte, la sección 3 describe el método utilizado para desarrollar los experimentos, la sección 4 contiene los resultados de los distintos autores en los cuatro grupos de n-gramas, la sección 5 contiene la discusión de los resultados y la sección 6 las conclusiones de la investigación.

2. Trabajos relacionados

A continuación, se describen una serie de trabajos relacionados con la detección de cambios de estilo de escritura, destacando las características estilométricas y la longitud de los textos (bloques de palabras) utilizados en los experimentos.

Patton y Can (2004) analizaron 4 novelas del autor turco Yasar Kemal, crearon bloques de 5000 palabras y utilizaron como características palabras frecuentes, conteo de sílabas, longitud de oraciones y de palabras, etiquetas POS, etc. Los resultados mostraron una clara separación entre los primeros trabajos y los últimos, las mejores características fueron palabras frecuentes y longitud de oraciones con 87% y 81% de precisión. En otra investigación Can y Patton (2004) realizaron experimentos con los autores turcos Cetin Altan y Yasar Kemal. El corpus de Kemal se formó con 2 novelas escritas en 1971 y 1998 mientras que el de Altan con artículos escritos de 1960 a 1969 y otros escritos en el año 2000; crearon bloques de 2,500 palabras y utilizaron como características la longitud de palabras, la riqueza de vocabulario y las palabras frecuentes. Patton y Can encontraron que la longitud de palabras fue mayor en los trabajos

recientes. Las pruebas de clasificación promediaron 92% de eficiencia, y concluyeron que tiempos de producción grandes entre novelas permiten una mejor separación entre los trabajos.

Hirst y Feng (2012) evaluaron el cambio de estilo de escritura en autores de los que se presume padecían Alzheimer; dividieron las novelas en bloques de 4,000 palabras y utilizaron como características palabras funcionales, signos de puntuación, n-gramas de caracter, etiquetas POS y entropía de POS. Hirst y Feng organizaron las novelas en 3 periodos (pero descartaron el periodo intermedio), definieron el número de novelas en cada periodo con base en los años en los que los autores fueron más productivos; se realizaron experimentos usando SVM como clasificador y validación cruzada. La precisión para cada uno de los autores fue de 81.7%, 77.5% y 55.8% respectivamente; los investigadores concluyeron que los estilos de los autores cambian a través del tiempo. Le (2010) investigó la presencia de demencia a través de cambios léxicos y sintácticos en autores literarios; creó muestras de novelas de 55,000 palabras; algunas características utilizadas fueron tamaño de vocabulario, repetición léxica, palabras de relleno, complejidad sintáctica y voz pasiva. Le encontró claras evidencias de una disminución lingüística en los últimos trabajos de Murdoch y Christie, mientras que P. D. James se mantuvo relativamente estable a lo largo de su carrera.

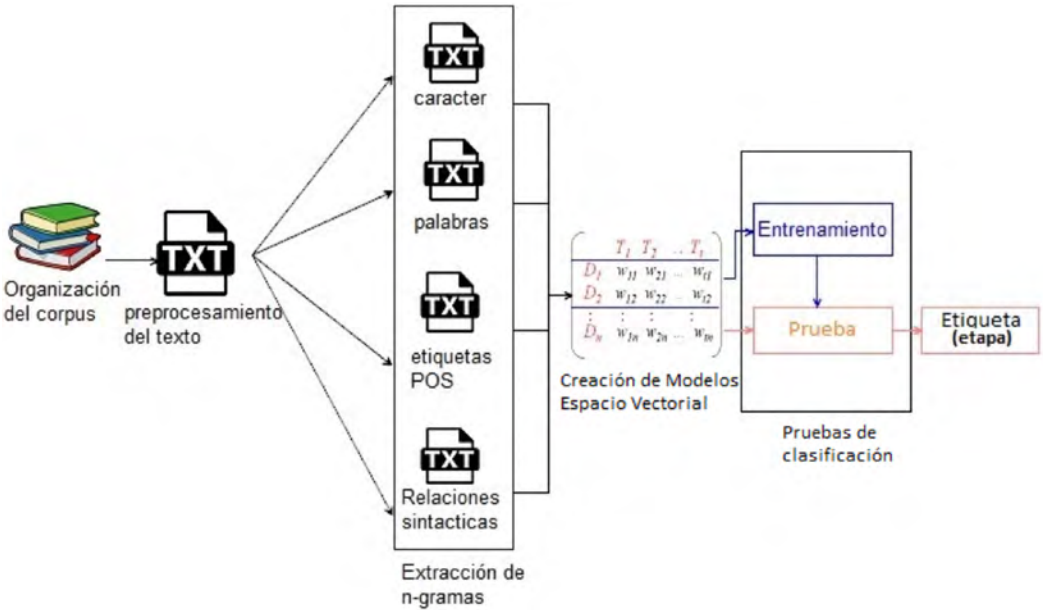
Lancashire y Hirst (2009) analizaron cambios en el vocabulario de la autora Agatha Christie, como indicios de la presencia de demencia; crearon bloques de 10,000 palabras (utilizaron los primeros 5 bloques) para obtener características como riqueza de vocabulario, n-gramas de palabras (colocaciones) y palabras "indefinidas". Lancashire y Hirst encontraron que la riqueza de vocabulario disminuyó conforme aumentaba la edad de la autora, las frases repetidas y el uso de palabras indefinidas mostraron el comportamiento contrario. Pennebaker y Stone (2003) analizaron novelas de distintos géneros literarios utilizando características LIWC (*linguistic inquiry and word count*), los resultados fueron correlacionados con la edad del autor al momento de escribir la obra; los investigadores concluyeron que la forma en que las personas usan el lenguaje cambia a lo largo de sus vidas y que muestran consistentes cambios en sus estilos lingüísticos en función de su edad. Pol (2005) creó un corpus de artículos de opinión en español que comprenden el periodo de un año; algunas características que utilizó fueron: riqueza de vocabulario, etiquetas POS, palabras funcionales y de contenido, longitud de oraciones, longitud de párrafos y estructuras sintácticas. Pol concluyó que los autores tienden a elegir las mismas opciones de la gran variedad que el lenguaje les ofrece. Spassova (2009) creó un corpus de autores de habla hispana, dividió los textos en muestras de 300 y 600 palabras para obtener las características bi-gramas y tri-gramas de etiquetas POS; organizó las novelas en 3 etapas (*aparente*, *intermedia* y *real*) y realizó experimentos evaluando tiempos de producción de 5 y 10 años entre novelas; Spassova identificó variación en novelas escritas con un mínimo de diferencia de 10 años. Williams y otros (2003) analizaron cartas escritas durante 20 años agrupadas en periodos de 2 años; utilizaron las diez últimas oraciones de cada carta para obtener los verbos (principales, subordinados y embebidos), la complejidad sintáctica, la longitud y cláusulas de cada oración. Los resultados mostraron

cambios importantes a partir de 1616, periodo en que el autor sufrió una enfermedad, posiblemente demencia vascular.

3. Metodología

Se propone un método para la detección de cambios de estilo a través del tiempo basado en métodos de aprendizaje supervisado. Dicho método utiliza una representación vectorial de los documentos usando como características cuatro tipos de n-gramas: de caracter, palabras, etiquetas POS y relaciones sintácticas (*syntactic relationship*, SR). La figura 1 muestra las etapas que conforman el proceso de la investigación. Los algoritmos de aprendizaje determinan a qué etapa pertenecen las muestras de un autor.

FIGURA 1
Metodología de los experimentos



A continuación, se describe con detalle cada sección de la metodología.

3.1. Organización del corpus

El corpus se formó con novelas descargadas del Proyecto Gutenberg[1]; se seleccionaron 7 autores nativos de habla inglesa con 9 novelas cada uno. Las novelas se organizaron con base

1 https://www.gutenberg.org/.

en la fecha de publicación y se definieron 3 etapas denominadas *inicial*, *media* y *final* con 3 novelas cada una. Entre dos etapas consecutivas existen por lo menos 2 años de separación entre las fechas de publicación de las novelas. La tabla 1 muestra la información obtenida del proceso anterior; la diferencia entre las etapas está ordenada en orden creciente por la columna *Inicial y Media*. La distribución de tiempos entre novelas permitirá averiguar cómo influye este parámetro en la detección de un cambio de estilo.

TABLA 1
Diferencia en años entre etapas

AUTOR	ETAPAS	
	INICIAL Y MEDIA	MEDIA Y FINAL
Louis Tracy	2	3
George Vaizey	6	4
Charles Dickens	7	6
Mark Twain	7	8
Frederick Marryat	8	5
George Macdonald	8	12
Booth Tarkington	9	3

3.2. Preprocesamiento del texto

Las novelas se convirtieron a letras minúsculas para evitar la redundancia a nivel de n-gramas de caracteres y palabras. Por ejemplo, la palabra *THE*, *The* y *tHe* se convirtió a *the* para obtener un único 3-grama de caracter; los signos de puntuación solo se utilizaron para crear n-gramas de caracter. Se extrajeron las oraciones de cada novela con el *software* NLTK[2] (*Natural Language ToolKit*); las oraciones de una y dos palabras se eliminaron debido a que, con excepción de los 3-gramas de caracter, los demás 3-gramas requieren de al menos 3 palabras. Con base en el número de oraciones, se crearon muestras de medias novelas (M) y novelas completas (C); las etiquetas POS se obtuvieron con el programa etiquetador de NLTK. Por último, el análisis sintáctico de las oraciones (*parsing*) para obtener el árbol sintáctico se realizó con la herramienta Stanford Parser[3].

2 http://www.nltk.org/.

3 https://nlp.stanford.edu/software/lex-parser.shtml.

3.3. Extracción de n-gramas

Después del preprocesamiento se obtuvieron los n-gramas de distintos tipos y las frecuencias de uso en muestras de medias novelas (M) y novelas completas (C). Se crearon n-gramas de caracter, POS y SR con **n=3**, ya que hay investigaciones que reportan resultados favorables con ese valor: en detección de plagio (Barrón-Cedeño y Rosso, 2009), en atribución de autoría (Escalante y otros, 2011; Sidorov y otros, 2014; Sapkota y otros, 2014), en categorización de textos (Addellatif y Zakaria, 2007) y en identificación de autores (Houvardas y Stamatatos, 2006).

Además, se crearon n-gramas de palabras con **n={1, 2, 3, 4}**. Los 3-gramas de caracter, palabras y POS se generaron con el programa *text2ngram*[4] (Velázquez, 2014) mientras que los 3-gramas SR con un script desarrollado en Python[5] (Sidorov y otros, 2014; Posadas-Durán y otros, 2015; Markov y otros, 2017).

Para reducir la cantidad de n-gramas se realizó una selección basada en frecuencia con un valor mínimo de 3; los n-gramas con frecuencia 1 y 2 no se utilizaron: provocan que el número de n-gramas aumente considerablemente. Entre menos frecuencia de uso tiene un n-grama, menor es la posibilidad de encontrarlo en un conjunto de textos (Stamatatos, 2006).

3.4. Creación del Modelo de Espacio Vectorial

Para los distintos grupos de n-gramas se crearon dos tipos de Modelo de Espacio Vectorial (*Vector Space Model*, VSM): en el primer escenario el VSM contiene n-gramas obtenidos con el criterio basado en la frecuencia de ocurrencia, en el segundo escenario se crearon VSM que contienen n-gramas seleccionados con el algoritmo de Análisis de Componentes Principales (*Principal Component Analysis*, PCA). La tabla 2 muestra los unigramas de palabras usando la frecuencia de ocurrencia como criterio de selección para muestras de medias novelas (M) y novelas completas (C).

El algoritmo PCA reduce las dimensiones de una matriz de tamaño ***m*** x ***n*** (un VSM), donde *m* es el número de muestras y *n* el número de n-gramas de la colección (las dimensiones). El algoritmo genera un nuevo VSM donde el número de dimensiones es **n=m**; las nuevas dimensiones reciben el nombre de *componentes principales*. El valor de **m** para muestras M y C fue de 18 y 9 respectivamente.

4 https://homepages.inf.ed.ac.uk/lzhang10/ngram.html.

5 http://www.cic.ipn.mx/~sidorov/MultiSNgrams_3_3.py.

TABLA 2
Número de unigramas por autor

AUTOR	M	C
BoothTarkington (BT)	5,441	6,704
CharlesDickens (CD)	12,411	14,432
FrederickMarryat (FM)	8,691	10,401
GeorgeMacdonald (GM)	6,763	8,312
GeorgeVaizey (GV)	5,350	6,621
LouisTracy (LT)	6,202	7,864
MarkTwain (MT)	10,062	11,994

Se utilizó el algoritmo PCA implementado en la herramienta scikit-learn[6] con un parámetro x=0.9 ($0 \geq \mathbf{x} \leq 1$), que representa el porcentaje de varianza acumulada que deben reunir los componentes principales. Se utilizó este porcentaje porque el último componente contiene una varianza que tiende a 0 y su contribución en la representación de los datos es mínima.

3.5. Pruebas de clasificación

Las pruebas de clasificación se realizaron utilizando algoritmos de aprendizaje automático supervisado como Máquinas de Vectores de Soporte (SVM), Naive Bayes Multinomial (MNB), Regresión Logística (LG) y Liblinear, debido a la capacidad que poseen para manejar modelos o vectores con gran número de dimensiones (Posadas-Durán, 2017). Se consideró como línea base un modelo aleatorio con 33% de exactitud que representa la probabilidad de clasificar de forma correcta un texto asignándolo de forma aleatoria a una de las 3 etapas. Con el fin de dar a conocer la configuración de los experimentos, la tabla 3 muestra las novelas para el análisis de estilo del autor BT.

A partir de esas novelas se obtuvo un total de 27 conjuntos de prueba con sus respectivos conjuntos de entrenamiento. Cada conjunto de prueba se formó con una novela de cada etapa (3 novelas) y con las dos novelas restantes se formó el conjunto de entrenamiento (6 novelas). Con estas combinaciones se asegura que todas las novelas sean parte de ambos conjuntos. La tabla 4 muestra los 27 conjuntos de prueba (CP) y las novelas que los conforman del autor BT generados con este proceso; los conjuntos de entrenamiento no se muestran por razones de espacio.

6 http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html.

TABLA 3
Organización de las novelas del autor BT

ETAPA 1		ETAPA 2		ETAPA 3	
NOVELA	**AÑO**	**NOVELA**	**AÑO**	**NOVELA**	**AÑO**
Gentleman	1899	Penrod	1914	Ramsey	1919
Vanrevels	1902	Turmoil	1915	AliceAdams	1921
Canaan	1905	Seventeen	1916	Julia	1922

TABLA 4
Conjuntos de prueba del autor BT

CP	NOVELAS	CP	NOVELAS	CP	NOVELAS
1	Gentleman, Penrod, Ramsey	10	Vanrevels, Penrod, Ramsey	19	Canaan, Penrod, Ramsey
2	Gentleman,Penrod, AliceAdams	11	Vanrevels, Penrod, AliceAdams	20	Canaan, Penrod, AliceAdams
3	Gentleman, Penrod, Julia	12	Vanrevels, Penrod, Julia	21	Canaan, Penrod, Julia
4	Gentleman, Turmoil, Ramsey	13	Vanrevels, Turmoil, Ramsey	22	Canaan, Turmoil, Ramsey
5	Gentleman, Turmoil, AliceAdams	14	Vanrevels, Turmoil, AliceAdams	23	Canaan, Turmoil, AliceAdams
6	Gentleman, Turmoil, Julia	15	Vanrevels, Turmoil, Julia	24	Canaan, Turmoil, Julia
7	Gentleman, Seventeen, Ramsey	16	Vanrevels, Seventeen, Ramsey	25	Canaan, Seventeen, Ramsey
8	Gentleman, Seventeen, AliceAdams	17	Vanrevels, Seventeen, AliceAdams	26	Canaan, Seventeen, AliceAdams
9	Gentleman, Seventeen, Julia	18	Vanrevels, Seventeen, Julia	27	Canaan, Seventeen, Julia

Las clases están balanceadas porque tienen el mismo número de muestras; esta condición hace factible el uso de la métrica exactitud (*accuracy*) (Garc, 2009). La exactitud se calcula mediante la fórmula 1:

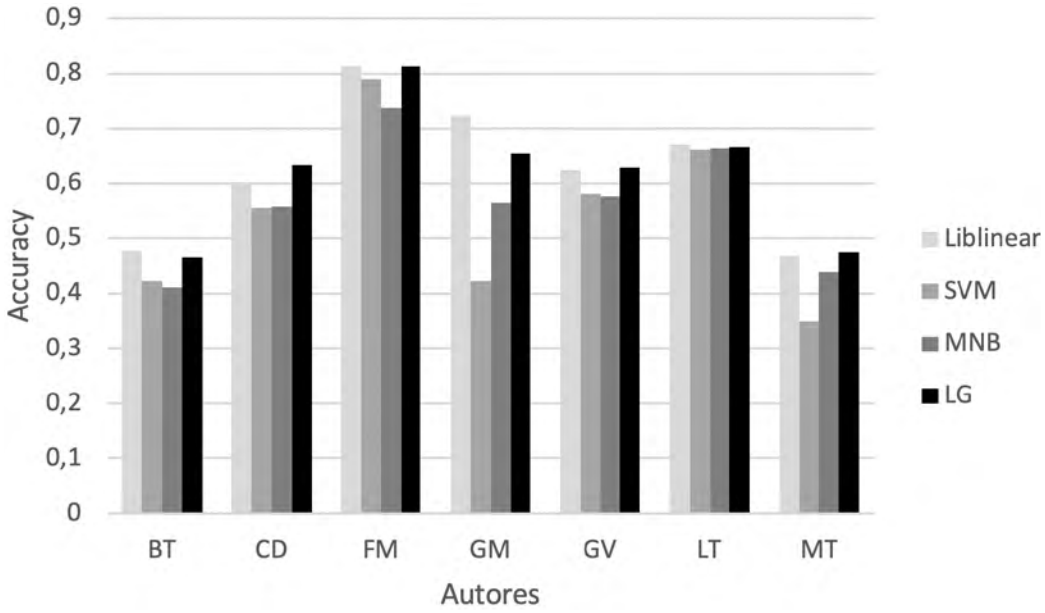
(1) $Accuracy=(tp+tn)/(tp+tn+fp+fn)$

donde **tp** significa *verdadero positivo*, **tn** *verdadero negativo*, **fp** *falso positivo* y **fn** *falso negativo*. Esta métrica mide la proporción de resultados verdaderos (tanto *tp* como *tn*) entre el número total de muestras evaluadas.

4. Resultados

La primera etapa del análisis de resultados consistió en evaluar cuál de los cuatro algoritmos de aprendizaje presentó el mejor promedio general de exactitud entre los cuatro grupos de n-gramas. La figura 2 muestra los resultados en trigramas utilizando muestras de novelas completas (C); los algoritmos Liblinear y *Logistic Regression* (LG) presentaron exactitud promedio muy similar; se optó por este último para continuar la presentación de resultados.

FIGURA 2
Actuación de los algoritmos entre los 4 grupos de 3-gramas en muestras C



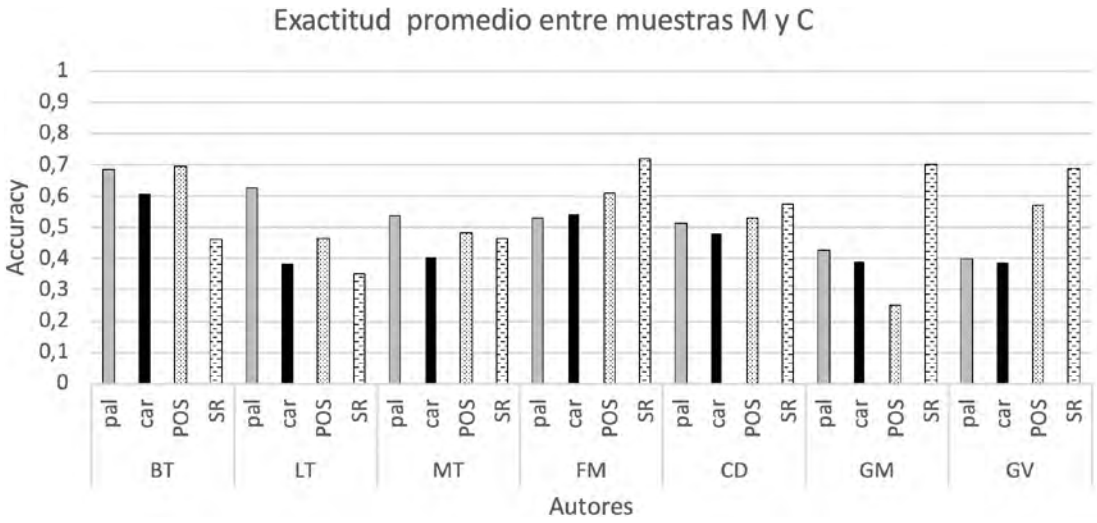
El orden de presentación de los resultados es el siguiente: promedios de los experimentos con el criterio de selección por frecuencia, promedios de los experimentos con selección con PCA, pruebas individuales de autores con alta exactitud, pruebas individuales de autores con baja exactitud y las pruebas estadísticas realizadas a cada autor.

4.1. Modelos sin selección de n-gramas

La figura 3 muestra la exactitud promedio de muestras de medias novelas (M) y novelas completas (C) de los distintos 3-gramas. Los autores están ordenados por 3-grama de palabras de forma decreciente. Recordemos que la línea base es 33%, todos los autores superan dicho porcentaje; sin embargo, la exactitud más alta apenas superó el 70% debido a la gran cantidad de n-gramas.

FIGURA 3

Exactitud con selección basada en frecuencia



4.2. Modelos con selección de n-gramas de palabras

La tabla 5 muestra la exactitud de los distintos tamaños de n-gramas de palabras utilizando el algoritmo PCA. La reducción de dimensiones generó un incremento en la exactidud al superar el 70% de los experimentos anteriores. Los autores CD, FM y LT obtuvieron una exactitud entre 70% y 89%, la más alta con repecto a la línea base, con uni-gramas de palabras o combinando todos los n-gramas.

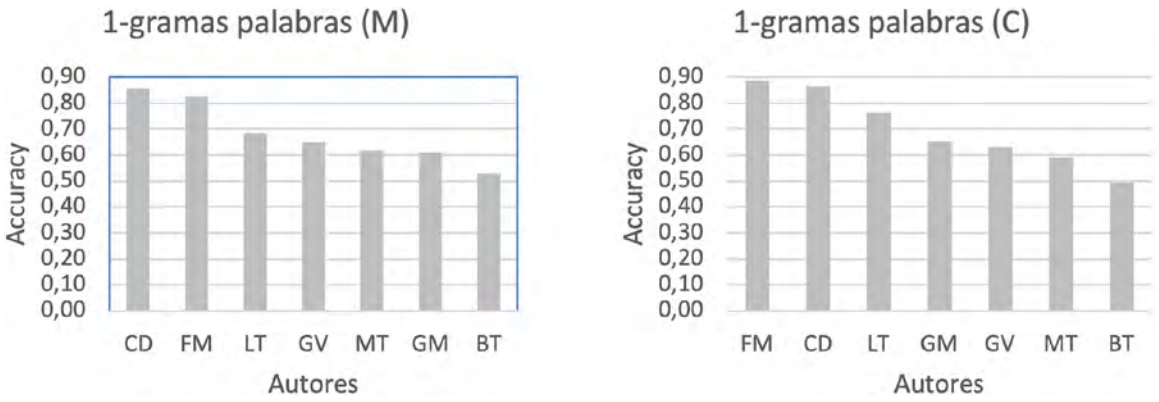
TABLA 5

Experimentos con n-gramas de palabras

ACCURACY (M=media C=completa)														
N	BT		CD		FM		GM		GV		LT		MT	
	M	C	M	C	M	C	M	C	M	C	M	C	M	C
1	0.53	0.49	0.86	0.86	0.83	0.89	0.61	0.65	0.65	0.63	0.69	0.77	0.62	0.59
2	0.62	0.54	0.80	0.83	0.72	0.84	0.64	0.64	0.56	0.56	0.78	0.84	0.59	0.58
3	0.57	0.53	0.72	0.75	0.61	0.72	0.56	0.51	0.39	0.44	0.59	0.70	0.64	0.63
4	0.36	0.25	0.72	0.75	0.80	0.80	0.43	0.49	0.48	0.89	0.56	0.53	0.52	0.51
1+2	0.57	0.53	0.83	0.81	0.80	0.88	0.65	0.63	0.59	0.62	0.75	0.80	0.64	0.60
1+3	0.54	0.52	0.83	0.78	0.80	0.89	0.64	0.60	0.62	0.58	0.71	0.77	0.66	0.53
1+4	0.53	0.49	0.87	0.83	0.83	0.89	0.62	0.65	0.64	0.65	0.69	0.77	0.60	0.57
2+3	0.62	0.56	0.78	0.81	0.69	0.81	0.63	0.65	0.54	0.48	0.78	0.84	0.67	0.62
2+4	0.61	0.54	0.80	0.84	0.70	0.84	0.63	0.65	0.56	0.57	0.78	0.84	0.64	0.57
3+4	0.57	0.52	0.71	0.73	0.63	0.74	0.58	0.53	0.44	0.51	0.62	0.70	0.66	0.67
1+2+3+4	0.57	0.53	0.83	0.81	0.80	0.88	0.65	0.63	0.59	0.62	0.75	0.80	0.64	0.60

En n-gramas donde n={2,3,4}, el mejor resultado se obtuvo con 2-gramas de palabras, la combinación de pares de n-gramas de distinta longitud no superó a unigramas ni a la combinación de los distintos tamaños. La figura 4 muestra la exactitud ordenada de forma decreciente en uni-gramas de palabras para muestras de medias novelas (M) y novelas completas (C), ambas obtuvieron resultados similares en la mayoría de los experimentos. En ambos casos los autores CD, FM, LT presentan exactitud alta y BT la exactitud más baja.

FIGURA 4
Exactitud en 1-gramas de palabras



4.3. Modelos con selección de 3-gramas caracter, POS y SR

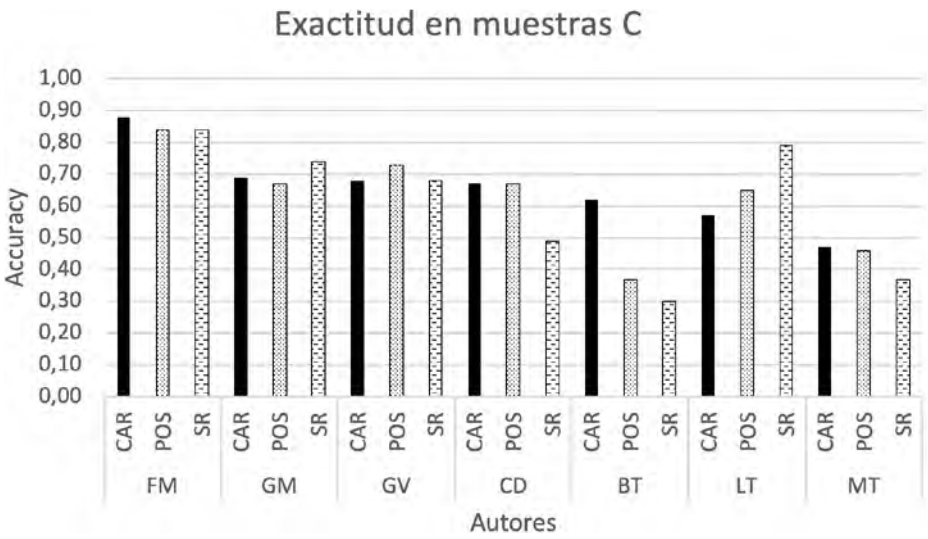
La tabla 6 muestra la exactitud promedio en muestras M y C en los distintos tipos de n-gramas y utilizando el algoritmo PCA. El autor FM obtuvo entre 78% y 88% de exactitud en los distintos 3-gramas, le siguen CD, GM, GV y LT con resultados entre 65% y 79% de exactitud y por último BT y MT con exactitud máxima de 62% en muestras C y 3-gramas de caracter.

TABLA 6
Experimentos con medias novelas y novelas completas

ACCURACY (M=medias C=completas)														
3-GRAMAS	BT		CD		FM		GM		GV		LT		MT	
	M	C	M	C	M	C	M	C	M	C	M	C	M	C
caracter	0.57	0.62	0.65	0.67	0.81	0.88	0.61	0.69	0.60	0.68	0.54	0.57	0.54	0.47
POS	0.46	0.37	0.69	0.67	0.78	0.84	0.66	0.67	0.65	0.73	0.62	0.65	0.49	0.46
SR	0.28	0.30	0.65	0.49	0.79	0.84	0.75	0.74	0.66	0.68	0.74	0.79	0.40	0.37

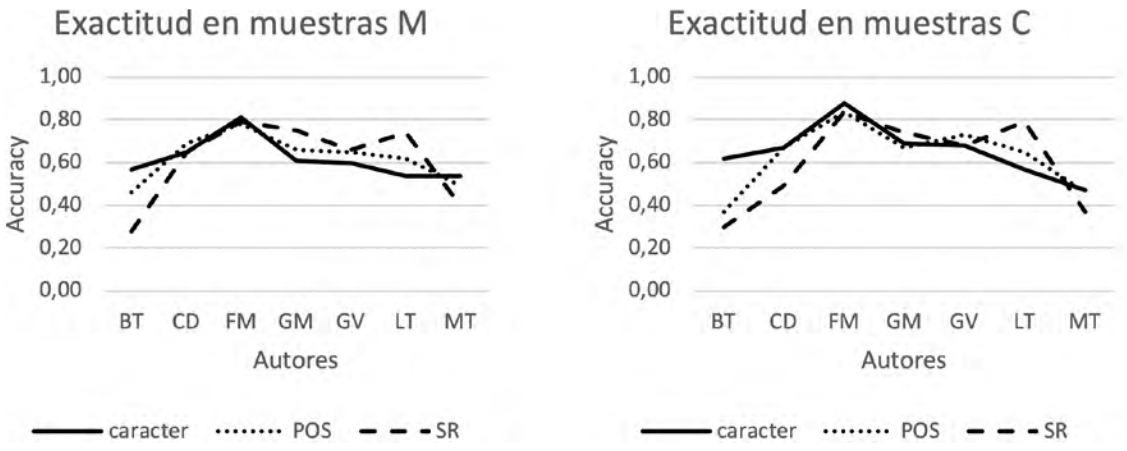
La figura 5 muestra con más detalle los resultados de cada autor en muestras C con la exactitud de 3-gramas de caracter ordenada en forma decreciente.

FIGURA 5
Exactitud ordenada por 3-gramas de caracter



La figura 6 muestra el tipo de n-grama que logró la exactitud en cada autor en ambas muestras. LT obtuvo 79% con 3-gramas SR, FM con valores que rondan el 80% de exactitud en los 3-grupos y BT con 3-gramas de caracter con aproximadamente 60% de exactitud.

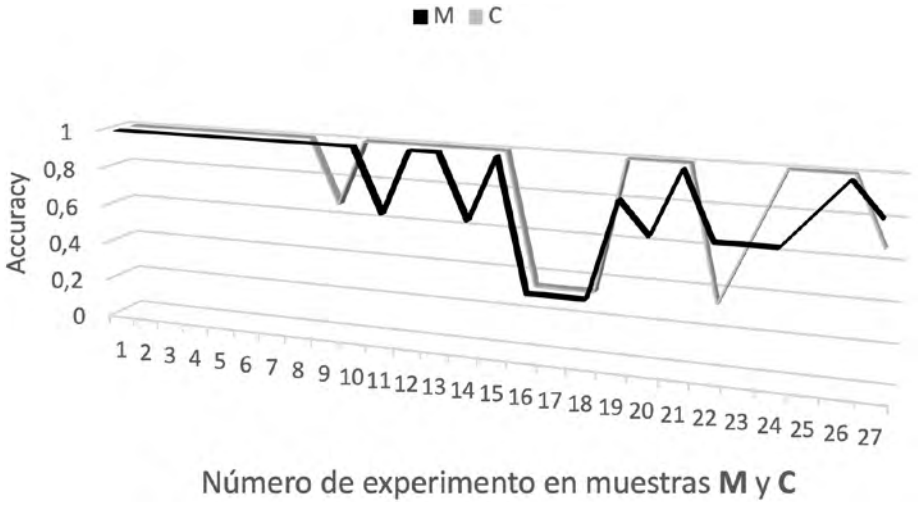
FIGURA 6
Variaciones entre 3-gramas en muestras M y C



4.4. Autores con mejor exactitud en uni-gramas de palabras

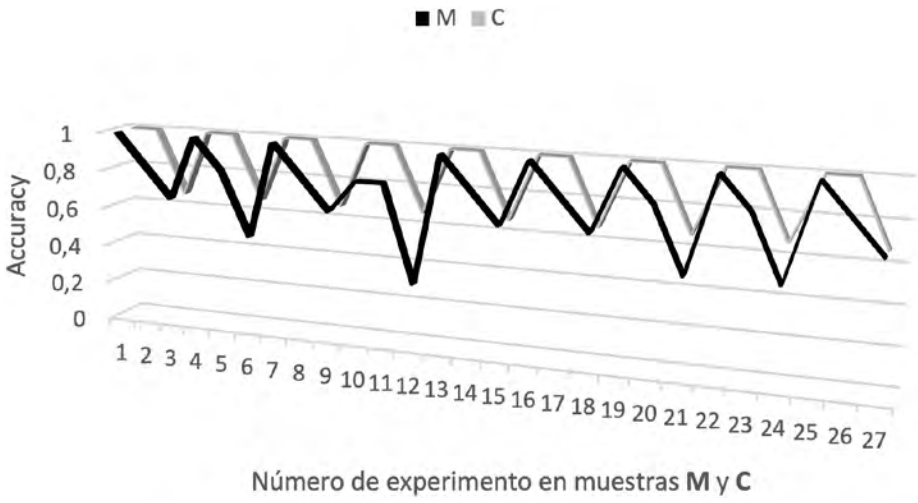
En esta sección se presenta la exactitud en cada uno de los 27 experimentos en muestras de medias novelas (M) y novelas completas (C). La figura 7 muestra a CD con 100% en el primer tercio de los experimentos; los experimentos 16, 17, 18 mostraron la exactitud minima de 33%.

FIGURA 7
Uni-gramas de palabras (CD)



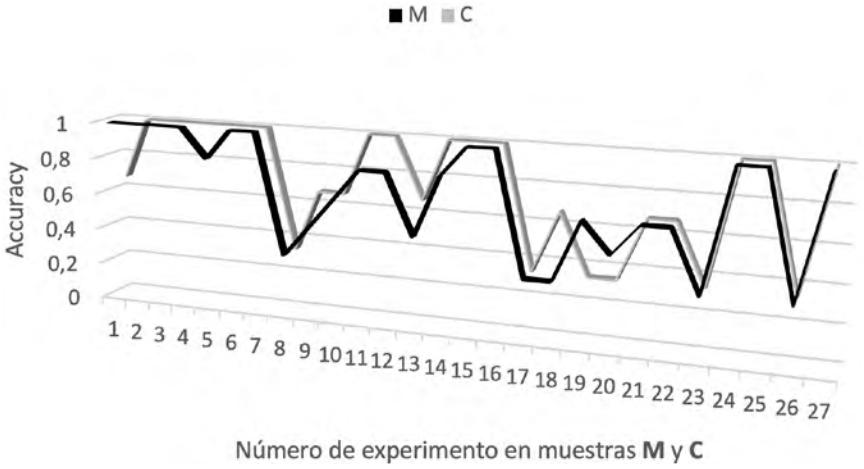
La figura 8 muestra a FM con mejores resultados en muestras de novelas completas (C). En la mayor parte de los experimentos obtiene entre 80% y 100% de exactitud. Las muestras M presentaron un comportamiento muy similar a muestras C, pero sin superarlas.

FIGURA 8
Uni-gramas de palabras (FM)



La figura 9 muestra que LT logra 100% de eficiencia en al menos un tercio de los experimentos. El resto de los experimentos supera el 66%; ambas muestras presentan un comportamiento bastante similar.

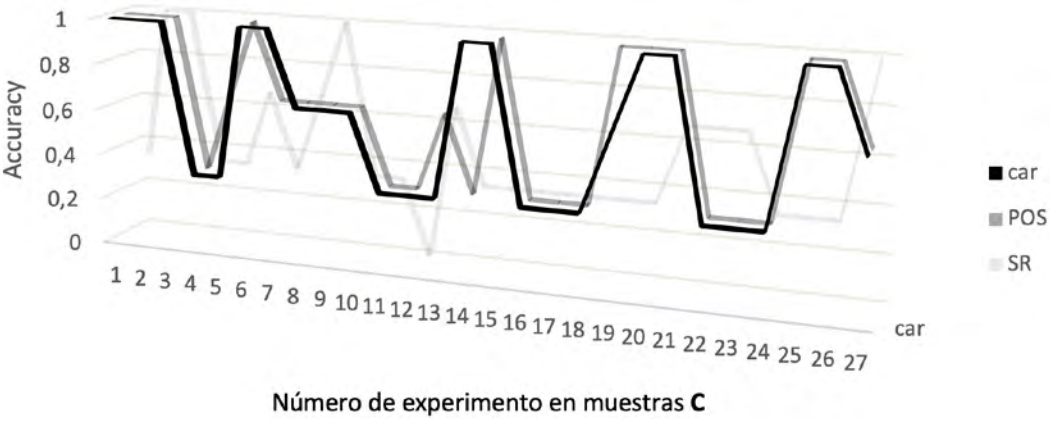
FIGURA 9
Uni-gramas de palabras (LT)



4.5. Autores con mejor exactitud en 3-gramas caracter, POS y SR

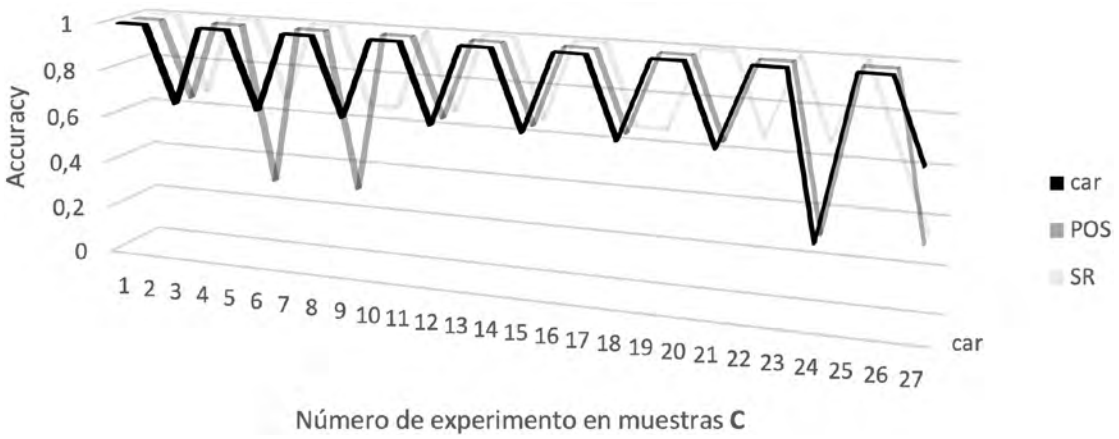
En esta sección se presenta la exactitud en cada uno de los 27 experimentos en muestras de novelas completas (C) utilizando 3-gramas de caracter, POS y SR. La figura 10 muestra a CD con una exactitud que supera el 70% en aproximadamente dos tercios de los experimentos. El experimento 13 tuvo exactitud de 0% en 3-gramas SR; en esa configuración, el algoritmo SVM no logró asignar a la clase correcta ninguna de las muestras.

FIGURA 10
Exactitud en novelas completas del autor CD



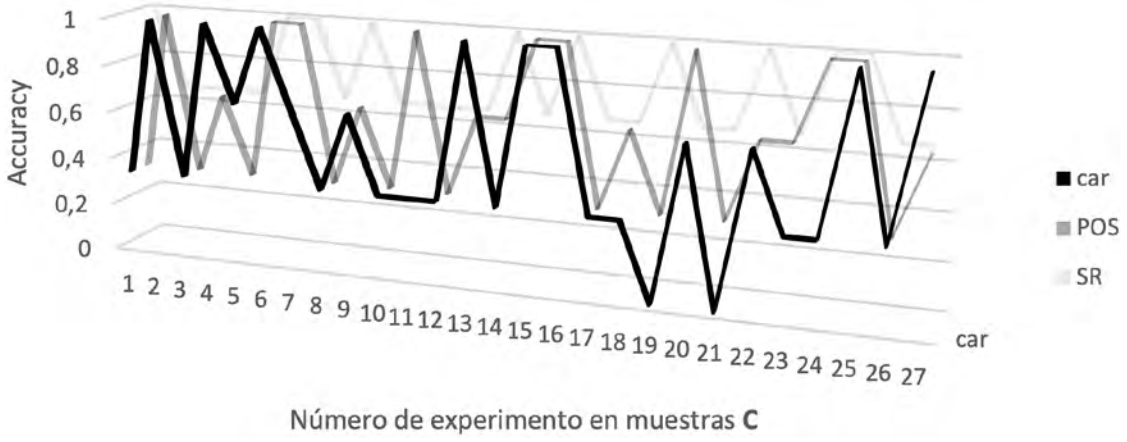
La figura 11 muestra que FM supera el 80% de exactitud en la mayor parte de los experimentos. Solo en 4 de los experimentos ocurrió la exactitud mínima de 33%. El comportamiento de los distintos grupos de 3-gramas fue muy similar.

FIGURA 11
Exactitud en novelas completas del autor FM



La figura 12 muestra a LT con una exactitud de entre 70% y 100% en dos tercios de los experimentos. El comportamiento de cada tipo de n-grama es irregular; sin embargo, los 3-gramas SR mostraron la mejor exactitud a lo largo de los experimentos. Se presentaron 2 experimentos con 0% de exactitud en 3-gramas de caracter.

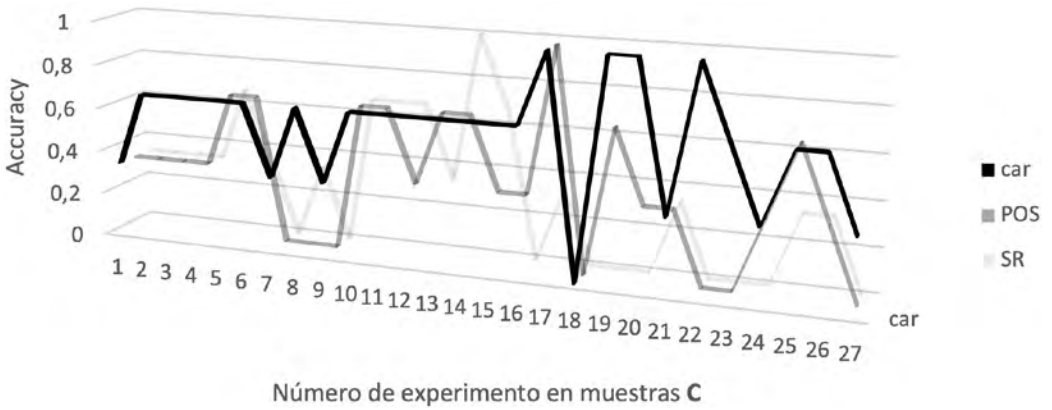
FIGURA 12
Exactitud en novelas completas del autor LT



4.6. Autores con baja exactitud en 3-gramas CAR, POS y SR

En esta sección se presenta la exactitud en cada uno de los 27 experimentos en muestras de novelas completas (C). La figura 13 muestra a BT con exactitud de 0% en prácticamente un tercio de los experimentos en 3-gramas POS y SR.

FIGURA 13
Exactitud en novelas completas del autor BT



La figura 14 muestra que MT obtiene exactitud 0% en los distintos grupos de 3-gramas; en general la mayor parte de los experimentos no superó el 66% de exactitud.

FIGURA 14
Exactitud en novelas completas del autor MT

4.7. Pruebas de significancia

Para verificar estadísticamente la hipótesis se utilizó la prueba *Wilcoxon Signed-Ranks* (Wilcoxon, 1945); la hipótesis se planteó de la siguiente manera:

Hipótesis nula (H_0): no existe cambio de estilo de escritura.
Hipótesis alternativa (H_1): existe cambio de estilo de escritura.

Demšar (2006) aporta dos argumentos a favor de este tipo de prueba: desde el punto de vista estadístico es más segura, ya que no asume distribuciones normales y que los valores atípi-

cos tienen menos efecto sobre la media. Demšar indica que esta prueba es una alternativa no paramétrica a la prueba *t* pareada que clasifica las diferencias en el rendimiento de dos clasificadores para cada conjunto de datos. Esta prueba requiere que los clasificadores comparados se evalúen utilizando las mismas muestras aleatorias y se realicen al menos cinco experimentos para cada método (Gómez-Adorno y otros, 2016); ambas condiciones se cumplen, por lo que es factible realizar la prueba. Un conjunto de datos representa la exactitud del algoritmo LG en cada uno de los 27 experimentos y el otro conjunto contiene los 27 resultados de un modelo aleatorio con 33% de exactitud, el cual representa la probabilidad de clasificar de forma correcta un texto asignándolo de forma aleatoria a una de las 3 etapas.

Se evaluaron los niveles de significancia $p<0.05$ y $p<0.01$. El primer nivel de significancia $p<0.05$ indica cambios significativos y el segundo $p<0.01$ indica cambios muy significativos; cuanto más pequeño es el valor de p, más evidencia existe para rechazar H_0.

Cabe recordar que los experimentos con n-gramas de palabras mostraron un promedio general de al menos 66% de exactitud (véase la tabla 5), indicando un cambio de estilo ya que más de la mitad de las muestras fueron clasificadas correctamente; esto era de esperarse, pero la detección de cambios en el nivel de palabras probablemente se debe a un cambio en el tópico de una novela. Sin embargo, los resultados con n-gramas de caracter, POS y SR mostraron indicios de que algunos autores no presentaban cambios. Para corroborar lo anterior, la prueba de hipótesis se realizó con estos 3-gramas, pero se incluyeron los 3-gramas de palabras para observar los resultados de las cuatro categorías.

La tabla 7 muestra los resultados de la prueba con $p<0.05$ en muestras de medias novelas (M) y novelas completas (C); la notación significa lo siguiente: **0** indica que se aceptó H_0 y **1** indica que se aceptó H_1. La prueba arrojó que BT no presentó cambios en 3-gramas SR en ambas muestras; con MT ocurrió algo similar, pero solo en muestras M; el resto de los autores presentó cambios significativos en el estilo de escritura.

TABLA 7
Prueba de significancia con $p<0.05$

SIGNIFICANCIA DE 3-GRAMAS								
AUTOR	M				C			
	CAR	PAL	POS	SR	CAR	PAL	POS	SR
BT	1	1	1	**0**	1	1	1	**0**
CD	1	1	1	1	1	1	1	1
FM	1	1	1	1	1	1	1	1
GM	1	1	1	1	1	1	1	1
GV	1	1	1	1	1	1	1	1
LT	1	1	1	1	1	1	1	1
MT	1	1	1	**0**	1	1	1	1

La tabla 8 muestra los resultados de la prueba con $p<0.01$ en muestras de medias novelas (M) y novelas completas (C). BT no presentó cambios en 3-gramas SR en ambas muestras, pero ahora la prueba con 3-gramas POS en muestras C indicó que no se detectaron cambios significativos. También el autor MT presentó resultados distintos a la prueba anterior; en esta ocasión los 3-gramas SR en muestras C indicaron que no se detectaron cambios. El resto de los autores presentan cambios muy significativos.

TABLA 8

Prueba de significancia con $p<0.01$

SIGNIFICANCIA DE 3-GRAMAS								
AUTOR	MEDIAS				COMPLETAS			
	CAR	PAL	POS	SR	CAR	PAL	POS	SR
BT	1	1	1	**0**	1	1	**0**	**0**
CD	1	1	1	1	1	1	1	1
FM	1	1	1	1	1	1	1	1
GM	1	1	1	1	1	1	1	1
GV	1	1	1	1	1	1	1	1
LT	1	1	1	1	1	1	1	1
MT	1	1	1	**0**	1	1	1	**0**

Los resultados de ambas pruebas indican que cinco autores (CD, FM, GM, GV y LT) presentaron cambios de significativos a muy significativos. Para complementar, se aplicó nuevamente la prueba de hipótesis en 4-gramas de palabras debido a que BT y MT presentaron la exactitud más baja en estos experimentos (véase la tabla 5). En la tabla 9 se observa que no se detectaron cambios para BT, mientras que con el resto de autores ocurrió lo contrario.

TABLA 9

Prueba de significancia con $p<0.01$ y $p<0.05$

4-GRAMAS DE PALABRAS				
AUTOR	M		C	
	0.01	0.05	0.01	0.05
BT	**0**	1	**0**	**0**
CD	1	1	1	1
FM	1	1	1	1
GM	1	1	1	1
GV	1	1	1	1
LT	1	1	1	1
MT	1	1	1	1

5. Discusión

Los escritores de obras literarias utilizan todos los recursos del lenguaje que su conocimiento les permite, con el objetivo de que el texto resultante tenga el efecto deseado en el lector. Los elementos más simples de manipular son las palabras, les siguen las categorías gramaticales y por último la información sintáctica.

Lo anterior coincide con los resultados obtenidos principalmente en uni-gramas de palabras. Los uni-gramas de palabras representan el vocabulario del autor. Es probable que el vocabulario utilizado entre una novela y otra cambie en función del tópico; de ser así, los cambios identificados no se deben a un cambio de estilo como tal, sino que se deben a un cambio de tópico. No se recomienda el uso de palabras para este tipo de estudios.

Un escritor experto dispone, además de las palabras, de otros recursos que están en otros niveles del lenguaje: el nivel sintáctico y semántico. En el nivel semántico existen recursos como las categorías de las palabras, la organización de las palabras por su significado (sinónimo, antónimo, homónimo, homófonos), catáforas, anáforas, etc. En el nivel sintáctico dispone de oraciones simples y compuestas, oraciones activas y pasivas, oraciones principales y subordinadas, etc. El grado de conocimiento de estos elementos permite al escritor decidir en qué momento utilizarlos.

A diferencia de las palabras, la manipulación consciente de este tipo de características puede resultar compleja. Sin embargo, los 3-gramas POS y SR mostraron resultados que indicaron cambios en 5 de los 7 autores. Cabe mencionar que los autores evaluados tuvieron largas y prolíficas carreras literarias y contaban con conocimientos profundos del lenguaje.

Por otro lado, los 3-gramas de caracter se crean en función de las palabras; pueden contener palabras formadas con 3 caracteres (*the*, *sky*, *bye*, etc.); en la mayoria de los casos contienen fragmentos de palabras y en un momento dado capturan signos de puntución.

6. Conclusiones

Se analizaron novelas de 7 autores de habla inglesa; dichas novelas se ordenaron de forma cronológica de acuerdo a la fecha de publicación y se definieron 3 etapas cada una con 3 novelas. Las novelas se dividieron en oraciones para crear muestras con medias novelas y novelas completas. Se obtuvieron n-gramas de distintas categorías para crear modelos VSM, los cuales se evaluaron con algoritmos de clasificación supervisados con el objetivo de identificar la etapa a la que pertenecen las muestras. La primera serie de experimentos demostró que el uso del algoritmo PCA mejoró la exactitud de los modelos en aproximadamente 10% en los distintos experimentos; las muestras de medias novelas (M) y novelas completas (C) presentaron resultados muy similares; sin embargo, se sugiere el uso de novelas completas.

La exactitud promedio de los autores CD, FM y LT estuvo entre 75% y 80% en los distintos tipos de n-gramas, lo que indica que en la mayoría de los 27 experimentos las muestras fueron clasificadas correctamente, lo que sugiere la presencia de cambios significativos en su estilo de escritura. En el otro extremo se encuentran los autores BT y MT, quienes en sus respectivos experimentos presentaron casos con 0% de exactitud, un indicador de que el clasificador no logró clasificar ninguna de las muestras de forma correcta, llevando a la conclusión preliminar de que no se detectaron cambios de estilo para estos autores.

Para corroborar lo anterior se realizó la prueba de hipótesis *Wilcoxon Signed Ranks*; dicha prueba confirmó que cinco autores (CD, FM, GM, GV y LT) presentaron cambios significativos y muy significativos en su estilo de escritura a nivel de caracteres, palabras, categorías gramaticales y estructuras sintácticas. Para los autores BT y LT la prueba mediante n-gramas de categorías gramaticales y n-gramas sintácticos no permitió identificar cambios significativos. De esta forma se concluye que algunos autores cambian su estilo de escritura y otros no.

Para profundizar en esta investigación, se contemplan como trabajos futuros: experimentos con dos etapas, la creación de un corpus formado con 100 autores, identificar estructuras gramaticales y sintácticas (típicas y atípicas) y el uso de toda la información relacionada a las palabras.

7. Bibliografía citada

Addellatif, Rahmoun, y Elberrichi Zakaria, 2007: "Experimenting N-grams in text categorization", *The International Arab Journals of Information Tecnology* 4, 377-385.

Argamon, Shlomo, Moshe Koppel, W. James Pennebaker y Jonathan Schler, 2009: "Automatically Profiling the Author of an Anonymous Text", *Communication of the ACM* 52 (2), 119-123.

Baayen, Harald, Hans van Halteren y Fiona Tweedie, 1996: "Outside the cave of shadows: using syntactic annotation to enhance authorship attribution", *Literary and Linguistic Computing* 11 (3), 121-132.

Barrón-Cedeño, Alberto, y Paolo Rosso, 2009: "Based on n-Grams Comparison", *European Conference on Information Retrieval*, 696-700.

Björklund, Johanna, y Niklas Zechner, 2016: "Syntactic methods for topic-independent authorship attribution", *Natural Language Engineering* 23 (5), 789-806 [disponible en https://doi.org/10.1017/S1351324917000249].

Brocardo, Marcelo Luis, Issa Traore, Sherif Saad e Isaac Woungang, 2013: "Authorship verification for short messages using stylometry" en *Proceedings of the International Conference on Computer, Information and Telecommunication Systems (CITS)*, 1-6.

Can, Fazli, y Jon M. Patton, 2004: "Change of writing style with time", *Computers and the Humanities* 38 (1), 61-82.

Coyotl-Morales, Rosa María, Luis Villaseñor Pineda, Manuel Montes-y-Gómez y Paolo Rosso, 2006: "Authorship Attribution Using Word Sequences", *Progress in Pattern Recognition, Image Analysis and Applications* 4225, 844-853 [disponible en https://doi.org/10.1007/11892755_87].

Demšar, Janez, 2006: "Statistical Comparisons of Classifiers over Multiple Data Sets", *Journal of Machine Learning Research* 7, 1-30.

Diederich, Joachim, Jörg Kindermann, Edda Leopold y Gerhard Paass, 2003: "Authorship attribution with support vector machines", *Applied Intelligence* 19 (1), 109-123.

Escalante, Hugo Jair, Thamar Solorio y Manuel Montes-y-Gómez, 2011: "Local Histograms of Character N-grams for Authorship Attribution" en *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, Portland, Oregon: Association for Computational Linguistics, 288-298.

Fuller, Simon, Phil Maguire y Philippe Moser, 2013: "A Deep Context Grammatical Model For Authorship Attribution" en *Proceedings of the 9th International Conference on Language Resources and Evaluation*, Reykjavik, Iceland: European Language Resources Association, 4488-4492.

Gamon, Michael, y Agnes Grey, 2004: "Linguistic correlates of style : authorship classification with deep linguistic analysis features" en *Proceedings of the 20th International Conference on Computational Linguistics*, 611-617.

García, Vicente, Ramón A. Mollineda y José Salvador Sánchez, 2009: "Index of Balanced Accuracy: A Performance Measure for Skewed Class Distributions" en Helder Araujo, Ana María Mendoça, Armando J. Pinho, María Inés Torres (eds.): *Pattern Recognition and Image Analysis*, Lecture Notes in Computer Science 5524, Berlin, Heidelberg: Springer, 441-448 [disponible en https://doi.org/10.1007/978-3-642-02172-5_57].

Gómez-Adorno, Helena, Ilia Markov, Grigori Sidorov, Juan-Pablo Posadas-Durán, Miguel A. Sánchez-Pérez y Liliana Chanona-Hernández, 2016: "Improving Feature Representation Based on a Neural Network for Author Profiling in Social Media Texts", *Computational Intelligence and Neuroscience* 2016.

Halteren, H. van, 2007: "Author verification by linguistic profiling", *ACM Transactions on Speech and Language Processing* 4 (1), 1-17.

Hirst, Graeme, y Olga Feiguina, 2007: "Bigrams of syntactic labels for authorship discrimination of short texts", *Literary and Linguistic Computing* 22 (4), 405-417.

Hirst, Graeme, y Vanessa Wei Feng, 2012: "Changes in Style in Authors with Alzheimer's Disease", *English Studies* 93 (3), 357-370.

Houvardas, John, y Efstathios Stamatatos, 2006: "N-gram feature selection for authorship identification", *Artificial Intelligence Methodology Systems and Applications* 4183, 77-86.

Kaster, Andreas, Stefan Siersdorfer y Gerhard Weikum, 2005: "Combining text and linguistic document representations for authorship attribution" en *SIGIR Workshop Stylistic Analysis of Text for Information Access STYLE*, 1(Pt 1), 27-35.

Kešelj, Vlado, Fuchun Peng, Nick Cercone y Calvin Thomas, 2003: "N-Gram-Based Author Profiles for Authorship Attribution" en *Pacific Association for Computational Linguistics*, 255-264.

Koppel, Moshe, y Jonathan Schler, 2003: "Exploiting Stylistic Idiosyncrasies for Authorship Attribution" en *IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, 69-72.

Kukushkina, Olga V., Anatoly A. Polikarpov y Dmitry V. Khmelev, 2001: "Using Literal and Grammatical Statistics for Authorship Attribution", *Problems of Information Transmission* 37 (2), 172-184 [disponible en https://doi.org/10.1023/A:1010478226705].

Lancashire, Ian, y Graeme Hirst, 2009: "Vocabulary Changes in Agatha Christie's Mysteries as an Indication of Dementia: A Case Study" en *9th Annual Rotman Research Institute Conference, Cognitive Aging: Research and Practice*, 1-5.

Le, Xuan, 2010: "Longitudinal Detection of Dementia Through Lexical and Syntactic Changes in Writing", *Literary and Linguistic Computing* 26 (4), 435-461 [disponible en https://doi.org/10.1093/llc/fqr013].

Luyckx, Kim, y Walter Daelemans, 2005: "Shallow Text Analysis and Machine Learning for Authorship Attribution" en *Proceedings of the 15th meeting of Computational Linguistics in the Netherlands (CLIN 2004)*, 149-160.

Luyckx, Kim, y Walter Daelemans, 2008a: "Authorship Attribution and Verification with Many Authors and Limited Data", *Computational Linguistics* (August), 513-520.

Luyckx, Kim, y Walter Daelemans, 2008b: "Using syntactic features to predict author personality from text" en *Proceedings of Digital Humanities*, 146-149.

Markov, Ilia, Lingzhen Chen, Carlo Strapparava y Grigori Sidorov, 2017: "CIC-FBK Approach to Native Language Identification" en *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, Copenhaheng: Association for Computational Linguistics, 374-381 [disponible en https://doi.org/10.18653/v1/W17-5042].

Patton, M. Jon, y Fazli Can, 2004: "A Detailed Stylometric Investigation of the Ince memed Tetralogy", *Computers and the Humanities* 38 (4), 457-467 [disponible en https://doi.org/10.1007/s10579-004-1906-6].

Peng, Fuchun, Dale Schuurmans, Vlado Keselj y Shaojun Wang, 2003: "Language independent authorship attribution using character level language models" en *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1*, 267-274.

Peng, Fuchun, Dale Schuurmans y Shaojun Wang, 2004: "Augmenting Naive Bayes Classifiers with Statistical Language Models", *Information Retrieval* 7 (3/4), 317-345.

Pennebaker, W. James, y Lori D. Stone, 2003: "Words of wisdom: language use over the life span", *Journal of Personality and Social Psychology* 85 (2), 291-301.

Pol, Marta Sanchez, 2005: "A Stylometry-Based Method to Measure Intra- and Inter-Authorial Faithfulness for Forensic Applications" en *28th Annual Inter-national ACM Conference on Research and Development in Information Retrieval*.

Posadas Durán, Juan-Pablo, Grigori Sidorov, Ildar Batyrshin y Elibeth Mirasol-Meléndez, 2015: "Author verification using syntactic N-grams" en *CEUR Workshop Proceedings*, 1391(Cic), 8-11.

Posadas Durán, Juan-Pablo, Ilia Markov, Helena Gómez-Adorno, Grigori Sidorov, Ildar Batyrshin, Alexander Gelbukh y Obdulia Pichardo-Lagunas, 2015: "Syntactic N-grams as features for the author profiling task" en *CEUR Workshop Proceedings*, 1391.

Posadas Durán, Juan Pablo Francisco, 2017: *Detección Automática de Plagio Usando Información Sintáctica* [http://tesis.ipn.mx/handle/123456789/21664].

Raghavan, Sindhu, Adriana Kovashka y Raymond Mooney, 2010: "Authorship Attribution Using Probabilistic Context-Free Grammars", *Computational Linguistics* 6, 38-42.

Rude, Stephanie, Eva-Maria Gortner y W. James Pennebaker, 2004: "Language use of depressed and depression-vulnerable college students", *Cognition & Emotion* 18 (8), 1121-1133.

Sanderson, Conrad, y Simon Guenter, 2006: "Short Text Authorship Attribution via Sequence Kernels, Markov Chains and Author Unmasking: An Investigation", *Computational Linguistics* (July), 482-491.

Sapkota, Upendra, Thamar Soloria, Manuel Montes-Y-Gómez, Steven Bethard y Paolo Rosso, 2014: "Cross-topic authorship attribution: Will out-of the topic data help?" *The 25th International Conference on Computational Linguistics (COLING 2014)*.

Sidorov, Grigori, Ildar Batyrshin y Juan-Pablo Posadas-Durán, 2014: "Complete Syntactic N-grams as Style Markers for Authorship Attribution" en *Human-Inspired Computing and Its Applications - 13th Mexican International Conference on Artificial Intelligence, MICAI 2014*, (Cic), 9-17.

Sidorov, Grigori, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh y Liliana Chanona-Hernández, 2012: "Syntactic Dependency-Based N-grams as Classification Features" en Ildar Batyrshin y Miguel González Mendoza (eds.): *Advances in Computational Intelligence*, Lecture Notes in Computer Science 7630, Berlin/Heidelberg: Springer, 1-11.

Spassova, Maria Stefanova, 2009: *El potencial discriminatorio de las secuencias de categorías gramaticales en la atribución forense de autoría de textos en español*, Universitat Pompeu Fabra (España).

Stamatatos, Efstathios, 2006: "Ensemble-based Author Identification Using Character N-grams", *ReCALL*, 41–46.

Turell, María Teresa, y Núria Gavaldà, 2013: "Towards an Index of Idiolectal Similitude (or Distance) in Forensic Authorship Analysis", *Journal of Law and Policy* XXI (2), 495-514.

Uzuner, Özlem, y Boris Katz, 2005: "A comparative study of language models for book and author recognition" en *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3651 LNAI, 969–980.

Castillo Velázquez, Francisco Antonio, 2014: *Un método automático para detección de autoría de texto usando información sintáctica*. Tesis de doctorado, Instituto Politécnico Nacional, México, D. F.

Wilcoxon, Frank, 1945: "Individual Comparisons by Ranking Methods", *Biometrics Bulletin* 1 (6), 80.

Williams, Kristine, Frederick Holmes, Susan Kemper y Janet Marquis, 2003: "Written language clues to cognitive changes of aging: an analysis of the letters of King James VI/I", *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences* 58 (1), P42-P44.

Yoon, Su-Yoon, y Suma Bhat, 2012: "Assessment of ESL learners' syntactic competence based on similarity measures" en *EMNLP-CoNLL 2012 - 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Proceedings of the Conference* (July), 600-608.

Zhao, Ying, y Justin Zobel, 2007: "Searching with style: Authorship attribution in classic literature", *Conferences in Research and Practice in Information Technology Series* 62, 59-68.