Características morfológicas para el análisis automático del estilo de escritura.

Morphological characteristics for the automatic analysis of the writing style.

Germán Ríos Toledo* (1).
Tecnológico Nacional de México campus Tuxtla Gutiérrez.
german.rt@tuxtla.tecnm.mx.

Cesar Alejandro Meza Pérez (2). Estudiante del Tecnológico Nacional de México campus Tuxtla Gutiérrez, L16270802@tuxtla.tecnm.mx.

Jonathan Velázquez Trinidad (3). Estudiante del Tecnológico Nacional de México campus Tuxtla Gutiérrez, L16270802@tuxtla.tecnm.mx.

Héctor Guerra Crespo (4). Tecnológico Nacional de México campus Tuxtla Gutiérrez, hector.gc@tuxtla.tecnm.mx.

Galdino Belizario Nango Solís (5). Tecnológico Nacional de México campus Tuxtla Gutiérrez, galdino.ns@tuxtla.tecnm.mx.

Aída Guillermina Cossio Martínez (6). Tecnológico Nacional de México campus Tuxtla Gutiérrez, aida.cm@tuxtla.tecnm.mx.

*corresponding author.

Artículo recibido en noviembre 08, 2021; aceptado en diciembre 14, 2021.

Resumen.

En el análisis automático del estilo de escritura se requieren marcadores de estilo que, idealmente sean persistentes a factores como el tópico del documento, ya sea que se traten de correos electrónicos, artículos, novelas, cartas. Otro factor importante a considerar es la cantidad de texto presente en dichos documentos. Comúnmente, un análisis de estilo de escritura se inicia utilizando las palabras como un marcador de estilo. Sin embargo, es relativamente sencillo manipular las palabras que una persona elige para elaborar sus escritos, además de que el vocabulario de un texto es dependiente del tópico que en él se trate. Por ello, es necesario explorar marcadores de estilo más robustos, principalmente aquellos relacionados con la estructura morfológica o sintáctica de las oraciones. En este artículo se evalúa el uso de n-gramas de etiquetas POS en el análisis del cambio de estilo de escritura en novelas de diez autores de habla inglesa. Las novelas de cada autor se ordenaron de cronológicamente para establecer dos etapas, a las cuales se les denominó etapas inicial y final. El problema se abordó por medio de un algoritmo de aprendizaje automático supervisado. Los resultados mostraron que en algunos de los autores evaluados existe evidencia clara de un cambio en el estilo de escritura a través del tiempo.

Palabras clave: Marcadores de estilo, n-gramas, etiquetas POS, frecuencia, características, clasificación

Abstract.

In the automatic analysis of the writing style, style markers are required that are ideally persistent to factors such as the topic of the document, whether they are e-mails, articles, novels, letters. Another important factor to consider is

the amount of text present in such documents. Commonly, a writing style analysis begins by using the words as a style marker. However, it is relatively easy to manipulate the words that a person chooses to elaborate their writings, in addition to the fact that the vocabulary of a text is dependent on the topic in question. Therefore, it is necessary to explore more robust style markers, mainly those related to the morphological or syntactic structure of sentences. This article evaluates the use of POS tag n-grams in the analysis of writing style change in novels by ten English-speaking authors. The novels of each author were ordered chronologically to establish two stages, which were called the initial and final stages. The problem was addressed using a supervised machine learning algorithm. The results showed that in some of the evaluated authors there is clear evidence of a change in writing style over time.

Keywords: Style markers, n-grams, POS tags, frequency, features, classification

1. Introducción.

Una forma común de representar objetos es utilizando las características que los definen y sus valores. Suponga que se desea describir un automóvil, algunas de las características que lo definen son marca, modelo, tipo de transmisión, número de cilindros, tipo de combustible, etc. Este tipo de representación también es aplicable a textos o documentos. En este caso las características son: nombre del autor, título, fecha de publicación, género literario, editorial, precio, etc. La cantidad de características necesarias depende del nivel de descripción de cada objeto. En el Procesamiento del Lenguaje Natural existen tareas como la detección del plagio, creación de perfiles de autor, protección del anonimato y atribución de autoría, todas ellas enfocadas al análisis automático del estilo de escritura. La idea detrás de este tipo de análisis consiste en que, por medio de un conjunto de características prestablecidas, sea posible identificar quién es el autor de un texto entre una colección de autores. En términos computacionales, el estilo de escritura se refiere a la frecuencia de uso de elementos del texto, conocidos como marcadores de estilo o características estilométricas (Toledo, Sánchez, Sidorov, & Durán, 2019), (en lo sucesivo se les denomina características).

Para cuantificar el estilo de escritura, es fundamental determinar qué características estilométricas se van a utilizar para representar los textos, como son palabras más frecuentes, tipos de palabras utilizadas (verbos, sustantivos, adjetivos) o longitud de la oración. A través de los años, se han propuesto muchas características para el análisis de estilo de escritura. (Stamatatos, 2009) detalla una clasificación de características. Otros factores que juegan un papel importante en el desempeño del análisis del estilo de escritura son: número de autores candidatos, número de textos por autor y cantidad de texto en cada documento. Actualmente, el principal desafío en el análisis de estilo de escritura es identificar las características adecuadas para cada tarea en particular del Procesamiento del Lenguaje Natural.

Una forma simple y natural de ver el texto es como una secuencia de elementos (palabras, dígitos, signos de puntuación) agrupados en oraciones. Dado que, el análisis de estilo de escritura se centra en la forma del texto y no en su contenido, es importante explorar la funcionalidad de las características estilométricas que sean robustas o independientes al tema que aborda el texto.

La atribución de autoría consiste en determinar de entre un conjunto de autores, quién es el autor de un texto en particular (Juola, 2008). Desde el enfoque de aprendizaje automático supervisado, esta tarea se propone como un problema de clasificación. Más precisamente como una tarea de categorización de texto multiclase y de etiqueta única (Sebastiani, 2002).

Este artículo se enfoca en el estudio de una característica estilométrica conocida como *n-grama*. Un n-grama es una secuencia de elementos obtenidos siguiendo el orden lineal del texto. Los elementos pueden ser caracteres, palabras, etiquetas POS e información sintáctica. Un n-grama supone una ventana imaginaria de tamaño n que mueve n elementos en cada iteración, hasta llegar al final del texto (Sidorov, 2013). La tabla 1 muestra diferentes tipos de n-gramas para la oración “Juan lee un libro interesante” y $n = \{1,2,3,4\}$. Para el ejemplo, se ha omitido el espacio en blanco.

Tabla 1. N-gramas para la oración "Juan lee un libro interesante".

n	**n-gramas de carácter**
1	J, u, a, n, l, e, e, u, n, l, i, b, r, o, i, n, t, e, r, e, s, a, n, t, e
2	Ju, ua, an, nl, le, ee, eu, un, nl, li, ib, br, ro, oi, in, nt, te, er, rs, sa, an, nt, te
3	Jua, anl, lee, eun, nli, ibr, roi, int, ter, rsa, ant
4	Juan, nlee, eunl, libr, roin, nter, resa, ante
n	**n-gramas de palabras**
1	Juan, lee, un, libro, interesante
2	Juan lee, lee un, un libro, libro interesante
3	Juan lee un, un libro interesante
4	Juan lee un libro
n	**n-gramas de etiquetas POS**
1	NP, VM, DI, NC, AQ
2	NP VM, VM DI, DI NC, NC AQ
3	NP VM DI, DI NC AQ
4	NP VM DI NC

A medida que aumenta el valor de n, la frecuencia de uso y la probabilidad de encontrar n-gramas comunes en diferentes documentos disminuye significativamente, estos problemas ocurren independientemente del tipo de n-grama. Según (Houvardas & Stamatatos, 2006), la selección de un valor óptimo de n depende del idioma.

2. Métodos y materiales.

En trabajos anteriores (Toledo, Sánchez, Sidorov, & Durán, 2019) y (Gómez-Adorno, Posadas-Durán, Ríos-Toledo, Sidorov, & Sierra, 2018) se analizó el cambio en el estilo de escritura de autores a través del tiempo. El corpus está compuesto por novelas de 10 autores de habla inglesa, cada autor cuenta con 6 novelas. Las novelas se obtuvieron del sitio web de Gutemberg Proyect[1]. El corpus se muestra en la Tabla 2.

Tabla 2. Corpus para el análisis del cambio de estilo de escritura.

Autor	**Etapa Inicial**		**Etapa Final**	
	Año	**Novela**	**Año**	**Novela**
Boot Tarkington(BT)	1899	Gentleman	1919	Ramsey
	1902	Vanrevels	1921	Alice Adams
	1905	Canaan	1922	Gentle Julia
Charles Dickens (CD)	1838	Nicholas Nickleby	1859	Two Cities
	1838	Oliver Twist	1861	Expectations
	1841	Barnaby	1865	Our mutual friend
Edgar Rice (ER)	1912	A Princess of Mars	1941	Llama of Gathol
	1914	The good of Mars	1942	Skeleton Men of Jupiter
	1918	The warlord of Mars	1944	Lasnd of terror
Frederick Marryat (FM)	1830	The King´s Own	1845	The Mission
	1831	Jacob Faithful	1847	New Forrest
	1831	Newton Forster	1848	The Little Savage
George Macdonald (GM)	1863	David Elginbrod	1888	Electrical Lady
	1864	Adela	1891	Flight of Shadow
	1865	Alec Forbes	1892	Hope of gospel

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

George Vaizey (GV)	1901	School Story	1914	Cassandra
	1902	Pixie	1914	College Girl
	1902	Houseful of Girls	1915	claire
Iris Murdoch (IM)	1954	Under the net	1973	The black prince
	1956	The Flight from the Enchanter	1975	a word child
	1958	The Bell	1995	Jackson's Dilema
John Buchan	1910	Prester John	1932	The gap in the curtain
	1915	The thirty nine steps	1936	the island of sheeps
	1916	Green mantle	1941	sick heart river
Louis Tracy (LT)	1903	Wings of morning	1912	Romance of NY
	1904	The revelers	1916	The day of wrath
	1905	Disapperance	1919	Mortimer Fenley
Mark Twain (MT)	1869	Innocents Abroad	1897	the Equator
	1872	Roughing It	1905	What is man?
	1876	Tom Sawyer	1906	Dollar

2.1. Preprocesamiento de texto.

Las novelas se dividieron en oraciones con NLTK[2] (Kit de herramientas de Lenguaje Natural). Las oraciones de 1 y 2 palabras se eliminaron, ya que los 3-gramas de palabras requieren al menos 3 términos o tokens. Cada novela se convirtió a caracteres en minúsculas para lograr que cadenas de caracteres superficialmente diferentes tengan la misma forma (por ejemplo: car, Car, cars y Cars se podría normalizar a car). Los signos de puntuación se utilizaron para formar parte de los n-gramas de palabras. Además, cada texto se etiquetó con el etiquetador POS de NLTK[3]. La tabla 3 muestra el número de oraciones para todo el corpus.

Tabla 3. Oraciones compuestas por 3 o más términos.

Autor	Fecha/Novela	Oraciones	Autor	Fecha/Novela	Oraciones
	1899 Gentleman	5,380	George Vaizey	1901 School	3,520
	1902 Vanrevels	2,802		1902 Houseful	4,128
Boot Tarkington	1905 Canaan	4,598		1903 Pixic	4,549
	1919 Ramsey	2,179		1914 Cassandra	7,194
	1921 Alice Adams	5,602		1914 College	4,840
	1922 Julia	4,312		1915 Claire	6,112
	1838 Nicholas	14,718	Iris Murdoch	1954 Under the net	1,702
	1838 Oliver	8,138		1956 The Flight from the Enchanter	2,029
Charles Dickens	1841 Barnaby	11,480		1958 The Bell	1,531
	1859 Two cities	7,007		1973 The Black Pince	3,446
	1860 Expectations	8,979		1975 A word child	4,356
	1805 Mutual Friend	18,539		1995 Jackson´s Dilema	1,689
	1912 A Princess of Mars	5,688		1910 Prester John	6,158
	1914 The gods of mars	7,265		1915 The thirty nine steps	3,404
	1918 The warlord of mars	5,291	John Buchan	1916 green mantle	8,244
Edgar Rice	1941 Llama of Gathol	7,274		1932 the gap in the curtain	6,838
	1942 Skeleton Men of Jupiter	1,883		1936 the island of sheeps	1,182
	1944 Land of Terror	1,431		1941 sick heart river	1,287
	1830 Kings Own	6,216		1903 Wings	6,234
	1832 Forster	5,944		1904 Revelers	6,480
Frederick Marryat	1834 Faithful	7,304	Louis Tracy	1905 Disappear	5,541
	1845 Mission	3,919		1912 Romance	4,106
	1847 New Forest	5,366		1916 Wrath	4,153
	1848 Little Savage	3,859		1919 Mottimer	4,894
	1863 Elginbrod	9,696		1869 Innocents	8,270
	1864 Adela Volt	2,988		1872 Roughing It	6,772
George Macdonald	1865 Forbes	10,315	Mark Twain	1876 Tom Sawyer	4,485

[2] https://www.nltk.org
[3] https://www.nltk.org/api/nltk.tag.html?highlight=stanford%20pos%20tagger

1888 Elect Lady	3,493	1897 Equator	8,411
1891 Flight Shadow	3,868	1906 Dollar	4,519
1892 Cospel	2,086	1906 What is man?	4,842

Es recomendable realizar la normalización de la longitud del texto para asegurar que todos los autores tengan aproximadamente la misma cantidad de información (Singhal, Salton, & Buckley, 1995). Sin embargo, este trabajo, el análisis del estilo de escritura es de tipo intra-autor. Un análisis de estilo de escritura intra-autor evalúa los escritos del mismo autor a través del tiempo. Para una situación más realista, el número de oraciones no está normalizado.

Las novelas de cada autor se dividieron en 4 bloques con el mismo número de oraciones. Así, el número de muestras de texto aumenta, pero la cantidad de texto o palabras en cada una de ellas disminuye. En la tabla 4 se observa el número de oraciones en cada muestra de las novelas de Booth Tarkington.

Tabla 4. Oraciones en las novelas de Booth Tarkington.

Novela	Número de ejemplos			
	1	2	3	4
Canaan	4,598	2,299	1,532	1,149
Gentleman	5,350	2,675	1,783	1,337
Penrod	3,841	1,740	1,160	870
Seventeen	3,917	1,958	1,305	979
Turmoil	5,892	2,946	1,964	1,473
Vanrevels	5,802	1,401	934	700

2.2. Características Estilométricas.

Se generaron n-gramas de palabras y etiquetas POS utilizando el programa $text2ngram$[4]. Al crear n-gramas, especifica la cardinalidad de n-grama y la frecuencia de corte. En la Sección 1 se mencionó que comúnmente, la cardinalidad de un n-grama es $n = \{1,2,3,4\}$. En trabajos previos, sobre tareas de procesamiento de lenguaje natural, tales como, detección de plagio, atribución de autoría, categorización de texto e identificación de autores, se reporta que $n = \{3\}$ proporciona el mejor rendimiento (Houvardas & Stamatatos, 2006), (Escalante, Solorio, & Montes, 2011), (Sapkota, Solorio, Montes, Bethard, & Rosso, 2014), (Sidorov, Velasquez, Stamatatos, Gelbukh, & Chanona-Hernández, 2014) y (Zuo, Zhao, & Banerjee, 2019). Además de determinar la cardinalidad de un n-grama, se debe encontrar el valor de frecuencia de corte apropiado. Esto depende principalmente de 4 factores: idioma, número de textos (muestras), cantidad de texto en cada muestra (en palabras u oraciones) y la característica estilométrica seleccionada para el análisis. La frecuencia indica la cantidad de veces que aparece una característica estilométrica en el texto. En general, cuanto más frecuente es un rasgo, más variación estilística captura (Stamatatos, 2009). Las tablas 5, 6 y 7 muestran el número de características (palabras n-gramas) para $n = \{1,2,3\}$.

Tabla 5. Total de 1-gramas de palabras.

Autor	1	2	3	4
BT	6,598	5,133	4,400	3,995
CD	17,346	14,017	12,357	11,274
ER	6,395	5,215	4,585	4,215
FM	10,173	8,133	7,164	6,562
GM	8,695	6,785	5,879	5,301
GV	6,971	5,316	4,463	3,966
IM	10,236	8,185	7,199	6,550
JB	7,581	6,010	5,277	4,794

[4] https://helpmanual.io/man1/text2ngram/

LT	7,681	5,783	4,921	4,335
MT	11,804	9,522	8,187	7,606

Tabla 6. Total de 2-gramas de palabras.

Autor	1	2	3	4
BT	8,850	6,187	4,966	4,279
CD	36,181	26,231	21,508	18,778
ER	10,170	7,354	6,090	5,326
FM	16,292	11,762	9,675	8,464
GM	13,672	9,693	7,951	6,828
GV	9,168	6,192	4,883	4,047
IM	18,908	13,483	11,013	9,588
JB	11,291	7,812	6,288	5,382
LT	8,749	5,837	4,582	3,850
MT	15,603	11,173	9,114	8,183

Tabla 7. Total de 3-gramas de palabras.

Autor	1	2	3	4
BT	2,613	1,569	1,134	949
CD	19,772	12,959	10,031	8,413
ER	4,554	2,984	2,293	1,899
FM	7,209	4,571	3,464	2,860
GM	5,046	3,084	2,299	1,868
GV	2,746	1,556	1,070	870
IM	8,649	5,426	4,103	3,455
JB	4,623	2,783	2,067	1,680
LT	2,427	1,366	961	717
MT	5,697	3,762	2,910	2,612

2.3. Matriz término-documento.

Los n-gramas y sus frecuencias se almacenan en la matriz denominada término-documento: las filas representan textos y las columnas el conjunto de n-gramas de un autor. La tabla 8, muestra una matriz término-documento de 1-gramas de palabras. Las matrices están ordenadas en orden descendente, la columna de la izquierda es el 1-grama que el autor usa con más frecuencia en sus textos.

Tabla 8. Las diez palabras más frecuentes de 1-grama para Booth Tarkington.

Ejemplos	the	and	to	of	a	he	in	was	his	i
Texto 1	2,757	1,342	993	1,146	1,057	803	680	568	552	371
Texto 2	2,663	1,630	1,249	1,170	1,162	701	709	598	548	483
Texto 3	1,890	828	750	880	700	452	523	512	341	144
Texto 4	1,460	739	708	569	531	428	375	290	325	259
Texto 5	1,921	869	932	989	807	646	516	442	475	318
Texto 6	2,043	1,038	996	917	744	666	509	467	484	381
Texto 7	903	597	534	512	404	377	288	265	236	179
Texto 8	624	528	535	371	329	195	218	206	111	282
Texto 9	1,313	866	1,150	803	829	469	471	375	286	448

Texto 10	1,177	860	1,177	643	731	714	389	344	328	565
Texto 11	1,388	965	912	797	752	601	489	452	335	335
Texto 12	1,107	749	871	680	572	453	384	340	253	376

2.4. Pruebas de clasificación.

A continuación se describen los conjuntos de entrenamiento y prueba. La Tabla 9 muestra las novelas de Booth Tarkington.

Tabla 9. Novelas de Booth Tarkington.

Etapa Inicial		**Etapa Final**	
Novela	**Año**	**Novela**	**Año**
Gentleman	1899	Ramsey	1919
Vanrevels	1902	Alice Adams	1921
Canaan	1905	Julia	1922

Los datos se dividieron en conjuntos de entrenamiento y prueba utilizando la estrategia Leave-One-Out: una novela por clase se usa una vez como conjunto de prueba y el resto para el conjunto de entrenamiento. Así, se crearon nueve tuplas de entrenamiento y prueba por autor. La tabla 10, muestra los conjuntos de prueba y entrenamiento para Booth Tarkington. En cada una de las nueve iteraciones, cada novelas solo pertenece a uno de los dos conjuntos.

Tabla 10. Conjuntos de prueba y entrenamiento de Booth Tarkington.

Conjunto de Prueba	**Conjunto de Entrenamiento**
Gentleman, Ramsey	VanRevels, Alice Adams, Canaan, Julia
Gentleman, Alice Adams	VanRevels, Ramsey, Canaan, Julia
Gentleman, Julia	VanRevels, Ramsey, Canaan, Alice Adams
VanRevels, Ramsey	Getnleman, Alice Adams, Canaan, Julia
VanRevels, Alice Adams	Getnleman, Ramsey, Canaan, Julia
VanRevels, Julia	Getnleman, Ramsey, Canaan, Ramsey
Canaan, Ramsey	Getnleman, Vanrevels, Alice Adams, Julia
Canaan, Alice Adams	Getnleman, Vanrevels, Ramsey, Julia
Canaan, Julia	Getnleman, Vanrevels, Ramsey, Alice Adams

La distribución de los conjuntos de prueba y entrenamiento se muestran en la tabla 11, una proporción de $1/3(\approx 33\%)$ para pruebas y $2/3$ $(\approx 67\%)$ para entrenamiento.

Tabla 11. Distribución para conjuntos de entrenamiento y prueba.

Tamaño	**Ejemplos**	**Ejemplos para prueba**	**Ejemplos para entrenamiento**
1	6	2	4
2	12	4	8
3	18	6	12
4	24	8	16
5	30	10	20

El problema se abordó como una atribución de autoría supervisada: dado un documento D y dos etapas $S = \{Inicial, Final\}$ para un autor único, determinar a cuál de las dos etapas en S, D pertenece. Este es un problema de clasificación binaria, la clase positiva está etiquetada con la etiqueta inicial. El clasificador binario predice las instancias del conjunto de pruebas como positivas o negativas y produce cuatro resultados: Verdadero Positivo (TP), Verdadero Negativo (TN), Falso Negativo (FN) y Falso Positivo (FP).

Las pruebas de clasificación se realizaron con el algoritmo de Aprendizaje Automático Supervisado Regresión Logística implementado en scikit-learn. La métrica de precisión no es una buena opción cuando hay el desbalance de clases. Sin embargo, en estos experimentos ambas clases están equilibradas (ver tabla 10), por lo que la precisión (accuracy) resulta apropiada para la evaluación (García, Mollineda, & Sánchez, 2009). La precisión es la fracción de predicciones que el modelo hizo correctamente. Su representación matemática se muestra en la siguiente ecuación.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Se obtuvo el promedio de nueve experimentos para los diferentes tamaños de texto. La línea de base es un modelo aleatorio hipotético con una 50% de precisión, la cual representa la probabilidad de asignar correctamente una muestra a una de las dos etapas.

3. Resultados.

3.1. n-gramas de palabras.

En las siguientes tablas, las etiquetas significan: **1h**= 100 n-gramas más frecuentes, **2h**= 200 n-gramas más frecuentes y así sucesivamente. La leyenda **todo** significa que se están utilizando todos los n-gramas del conjunto.

La tabla 12 muestra el total de n-gramas de palabras para $n = \{1,2,3,4\}$. A medida que el texto se divide en bloques más pequeños, la cantidad de características tiende a disminuir. Por ejemplo, para el autor ER en 2-gramas de palabras pasa de 10,170 en novela completa (1) a 5,326 en cuartos de novela (4).

Tabla 12. Total de características de n-gramas de palabras.

Autor	1-gramas de palabras				2-gramas de palabras				3- gramas de palabras			
	1	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
BT	6,598	5,133	4,400	3,995	8,850	6,187	4,966	4,279	2,613	1,569	1,134	949
CD	17,346	14,017	12,357	11,274	36,181	26,231	21,508	18,778	19,772	12,959	10,031	8,413
ER	6,395	5,215	4,585	4,215	10,170	7,354	6,090	5,326	4,554	2,984	2,293	1,899
FM	10,173	8,133	7,164	6,562	16,292	11,762	9,675	8,464	7,209	4,571	3,464	2,860
GM	8,695	6,785	5,879	5,301	13,672	9,693	7,951	6,828	5,046	3,084	2,299	1,868
GV	6,971	5,316	4,463	3,966	9,168	6,192	4,883	4,047	2,746	1,556	1,070	870
IM	10,236	8,185	7,199	6,550	18,908	13,483	11,013	9,588	8,649	5,426	4,103	3,455
JB	7,581	6,010	5,277	4,794	11,291	7,812	6,288	5,382	4,623	2,783	2,067	1,680
LT	7,681	5,783	4,921	4,335	8,749	5,837	4,582	3,850	2,427	1,366	961	717
MT	11,804	9,522	8,187	7,606	15,603	11,173	9,114	8,183	5,697	3,762	2,910	2,612

La tabla 13, muestra la precisión con 1-gramas de palabras. La precisión más alta es para los autores BT, ER y JB que promedian al menos el 95%, seguidos de LT y MT con resultados que oscilan entre el 73 y el 82%. Los autores restantes apenas logran un 68% de precisión. En este sentido, GM es el autor con menor precisión con 50% en novelas completas (1).

Tabla 13. Precisión con 1-grama de palabras.

Autor	1					2					3					4				
	1h	**2h**	**3h**	**4h**	**Todo**	**1h**	**2h**	**3h**	**4h**	**Todo**	**1h**	**2h**	**3h**	**4h**	**Todo**	**1h**	**2h**	**3h**	**4h**	**Todo**
BT	100	100	100	100	100	100	100	97	100	94	100	98	98	91	94	90	86	83	83	96
CD	67	67	67	67	61	61	61	61	61	61	63	63	63	63	63	64	65	65	65	65
ER	100	100	100	100	100	94	94	92	92	94	96	96	98	94	98	96	96	96	96	94
FM	56	56	56	56	56	58	58	58	58	58	61	61	65	63	61	64	64	65	65	65
GM	44	50	50	50	50	64	64	64	67	64	67	69	70	72	70	61	61	58	60	62
GV	67	72	67	67	67	78	96	67	67	64	74	74	70	69	70	61	68	67	71	68
IM	67	67	67	67	67	64	67	67	67	67	67	67	63	67	65	68	68	68	68	68
JB	100	100	100	100	100	100	97	97	100	97	96	96	94	96	96	93	93	93	93	93
LT	83	83	83	83	83	92	92	92	89	86	89	89	89	89	87	90	90	90	89	89
MT	72	72	72	72	67	75	75	75	75	75	67	69	67	67	67	79	79	79	78	74

La tabla 14, muestra la precisión con 2-gramas de palabras. Al igual que en experimentos anteriores, los autores BT, ER y JB logran hasta un 100% de precisión en novelas completas (1), así como en medias novelas (2). Contrariamente a 1-gramas de palabras, la precisión de los autores CD, GV e IM mejora para superar el 70%, incluso el 90% en tercios (3) y cuartos (4) de novelas. Con 2-gramas de palabras, el autor GM muestra la precisión más baja en los diferentes tamaños de novelas e independientemente del número de características utilizadas. Se destaca que la precisión del autor de LT disminuye de manera significativa a una media del 72% en los diferentes tamaños de novelas, resultado que dista mucho de 1-gramas de palabras de 90%. Por el contrario, el autor MT, mejoró su precisión en novelas completas (1), pasando del 72% en 1-gramas de palabras a 83% en 2-gramas de palabras. En los tamaños restantes de la novela, este autor también mostró una disminución en la precisión. Contrariamente a lo esperado, usar todas las características (**todo**) no garantiza una mejor precisión. Todos los autores muestran una disminución al usar todas las características con respecto a los n-gramas más frecuentes.

Tabla 14. Precisión con 2-gramas de palabras.

Autor	1					2					3					4				
	1h	**2h**	**3h**	**4h**	**Todo**	**1h**	**2h**	**3h**	**4h**	**Todo**	**1h**	**2h**	**3h**	**4h**	**Todo**	**1h**	**2h**	**3h**	**4h**	**Todo**
BT	100	100	100	100	100	94	97	97	97	86	93	94	94	94	89	96	96	94	93	89
CD	83	83	83	83	83	83	83	83	83	83	72	67	69	61	83	67	72	68	65	79
ER	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	99	100	100	99
FM	56	61	61	61	56	58	64	58	58	53	69	70	63	63	61	72	69	71	72	70
GM	50	56	50	56	56	47	50	50	50	53	48	52	54	54	54	51	53	54	54	54
GV	56	72	61	61	50	92	92	92	92	92	96	96	98	98	93	96	93	94	93	90
IM	72	78	78	78	67	61	72	75	75	75	65	67	69	69	72	64	69	71	69	67
JB	89	100	100	100	100	100	100	100	100	100	93	94	94	94	94	94	92	94	92	96
LT	78	72	72	72	67	67	67	64	64	67	80	74	76	76	72	81	72	74	74	74
MT	83	83	83	83	83	67	67	67	67	67	56	57	57	63	56	60	61	62	57	58

La tabla 15 muestra la precisión usando 3-gramas de palabras. La precisión disminuyó con respecto 1-gramas y 2-gramas de palabras (Ver tablas 13 y 14). La precisión de BT disminuyó aproximadamente un 10% en tercios de novelas (3) y hasta un 20% en cuartos de novela (4). En general, para el autor ER la precisión disminuyó un 20%, además de que en cuartos de novela (4) no superó el 70%. Sin embargo, la precisión del autor JB se mantuvo constante en todos los tamaños de novela, con una precisión del 95 al 100%. Así mismo, el autor GV mostró resultados constantes cercanos al 80%. Un valor atípico es el autor LT: la precisión con palabras de 3-gramas es superior a las palabras de 2-gramas hasta en un 10% con novelas completas (1) y medias novelas (2). Por otro lado, la precisión con 3-gramas de palabras es menor que la de 2-gramas de palabras hasta en un 20% en tercios y cuartos de novelas.

Tabla 15. Precisión con 3-gramas de palabras.

Autor	1					2					3					4				
	1h	2h	3h	4h	Todo	1h	2h	3h	4h	Todo	1h	2h	3h	4h	Todo	1h	2h	3h	4h	Todo
BT	100	94	94	89	83	94	94	92	92	81	83	83	87	85	81	75	76	76	75	72
CD	61	61	61	67	89	56	61	64	64	75	57	61	67	65	69	56	67	65	69	71
ER	50	72	83	83	83	64	69	69	67	72	67	72	69	70	67	67	67	67	64	64
FM	78	72	67	72	78	69	67	69	69	75	69	67	70	70	67	65	65	65	65	64
GM	67	72	72	72	61	64	64	64	58	58	63	59	67	56	57	57	57	56	57	61
GV	83	83	78	78	83	78	78	78	78	81	80	80	80	80	78	78	81	81	79	79
IM	72	67	67	67	72	78	78	75	75	69	76	76	76	76	69	71	75	74	72	71
JB	100	100	100	100	100	100	100	100	100	100	96	96	94	94	96	96	94	94	94	94
LT	94	89	83	78	56	78	75	67	58	53	56	61	57	57	59	60	56	57	57	54
MT	67	61	61	61	67	56	53	53	53	56	52	56	56	54	54	53	56	57	57	56

3.2. n-gramas de etiquetas POS

La tabla 16, muestra el número total de n-gramas de etiquetas POS para $n = \{1,2,3,4\}$. El número de características disminuye a medida que el texto se divide en bloques más pequeños. Cabe hacer notar que en 1-gramas de etiquetas POS hay solo 33 o 34 características, ya que son las que conforman el conjunto estándar de etiquetas POS: CC=Conjunción coordinadora, CD=número cardinal, DT= determinante, JJ=adjetivo, etc. Para conocer el número de etiquetas y su significado, consulte el proyecto Penn Treebank[5]

Tabla 16. Total de características de n-gramas de etiquetas POS.

Autor	POS 1 - gramas				POS 2 - gramas				POS 3 - gramas			
	1	2	3	4	1	2	3	4	1	2	3	4
BT	33	33	33	33	736	668	640	617	4,606	3,844	3,412	3,147
CD	34	34	34	34	859	823	793	775	7,629	6,646	6,100	5,691
ER	34	34	34	34	658	611	578	567	3,983	3,356	2,984	2,767
FM	33	33	33	33	743	702	680	663	5,612	4,795	4,327	4,031
GM	33	33	33	33	748	712	695	661	5,428	4,540	4,134	3,853
GV	33	33	33	33	697	647	616	602	4,594	3,824	3,431	3,145
IM	35	35	35	35	756	714	693	674	5,719	4,889	4,428	4,148
JB	34	34	34	34	691	636	598	572	4,353	3,679	3,277	3,039
LT	34	34	34	34	693	644	620	597	4,265	3,541	3,204	2,898
MT	34	34	34	34	735	704	674	655	5,607	4,859	4,410	4,157

Debido al reducido número de características en 1-gramas de etiquetas POS, solo se realizaron experimentos con todas las características. La tabla 17, muestra que los autores ER y JB logran una precisión de hasta el 100%, seguidos por los autores BT, FM e IM con una precisión que va del 80 al 90%. Por otro lado, el autor CD apenas alcanza el 61% de precisión en novelas completas (1) y medias novelas (2).

[5] https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Tabla 17. Precisión con 1-gramas de etiquetas POS y todas las características.

Autor	1	2	3	4
BT	86	94	87	96
CD	61	61	70	68
ER	100	100	100	100
FM	89	92	89	92
GM	72	50	41	44
GV	67	72	69	71
IM	89	78	81	78
JB	100	100	93	93
LT	72	75	70	67
MT	67	78	72	71

La tabla 18 muestra la precisión con 2-gramas de etiquetas POS. Se observa una tendencia similar a los experimentos con 1-gramas de etiquetas POS: los autores ER, JB, BT e IM son los mejores clasificados, seguidos de CD, FM y MT. Cabe señalar que los autores de GM y LT tienen la precisión más baja, independientemente del número de características y tamaños de la muestra.

Tabla 18. Precisión con 2-gramas de etiquetas POS.

Autor	**1**					**2**					**3**					**4**				
	1h	**2h**	**3h**	**4h**	**Todo**	**1h**	**2h**	**3h**	**4h**	**Todo**	**1h**	**2h**	**3h**	**4h**	**Todo**	**1h**	**2h**	**3h**	**4h**	**Todo**
BT	89	89	89	89	89	97	97	100	100	97	94	94	94	94	94	92	90	92	93	88
CD	72	72	72	72	72	75	75	75	75	78	85	83	83	83	83	76	76	75	75	75
ER	100	100	100	100	100	100	100	100	97	100	94	98	96	98	100	94	96	96	96	96
FM	78	78	78	78	78	83	83	83	83	83	83	83	83	83	83	85	85	85	85	85
GM	67	67	67	67	67	42	42	42	42	42	50	52	52	52	52	44	43	43	43	43
GV	61	61	61	61	61	69	69	69	69	69	70	69	69	69	69	72	74	74	74	74
IM	83	83	83	83	83	69	69	69	69	69	78	78	78	78	78	72	72	72	72	72
JB	100	100	100	100	100	100	100	100	100	100	91	91	89	91	91	96	93	93	93	93
LT	56	56	56	56	56	67	67	67	67	67	65	65	65	65	65	71	69	69	69	69
MT	72	72	72	72	72	78	78	78	78	78	70	72	70	70	70	67	68	68	68	68

La tabla 19 muestra la precisión con 3-gramas de etiquetas POS. Persiste la tendencia en todos los autores: El autor GM es el peor clasificado ya que en promedio no supera el 55% de precisión. Sin embargo, en estos experimentos el autor LT supera el 70% al utilizar novelas completas (1) y medias novelas (2). Otro punto a destacar es que el autor GV pasó de 74% en 2-gramas de etiquetas POS a un máximo de 92% en cuartos de novelas (4) 3-gramas de etiquetas POS.

Tabla 19. Precisión con 3-gramas de etiquetas POS.

Autor	**1**					**2**					**3**					**4**				
	1h	**2h**	**3h**	**4h**	**Todo**	**1h**	**2h**	**3h**	**4h**	**Todo**	**1h**	**2h**	**3h**	**4h**	**Todo**	**1h**	**2h**	**3h**	**4h**	**Todo**
BT	89	94	100	100	100	89	92	97	97	94	87	89	89	89	89	81	81	86	86	85
CD	78	72	78	72	72	92	94	94	92	89	94	96	93	91	91	93	93	92	92	88
ER	100	100	100	100	100	92	92	94	92	97	94	94	94	94	94	94	93	93	93	93
FM	78	78	78	78	78	81	81	81	81	81	83	83	83	83	81	82	82	82	83	83
GM	50	56	56	56	61	47	47	47	50	44	50	54	59	54	57	50	51	49	56	57
GV	61	61	61	61	61	67	67	67	67	67	85	74	70	80	72	90	92	90	83	83
IM	89	89	89	89	89	69	72	72	72	72	76	78	76	74	76	76	74	74	74	74
JB	100	100	100	100	100	92	94	94	94	100	93	85	85	89	81	90	89	92	90	89
LT	72	78	67	67	67	72	75	72	72	72	59	67	67	63	65	64	69	62	74	69
MT	83	83	78	78	78	83	78	81	78	78	76	76	76	69	74	67	68	72	72	65

Discusión.

La forma natural de evaluar el estilo de escritura de un autor es con las palabras que utiliza. Dichas palabras conforman el vocabulario del autor. De esta forma se conoce la frecuencia de uso: aquellas que usa con más frecuencia, o con menos frecuencia, palabras que usa una sola vez o dos veces, entre otras. Sin embargo, un análisis de estilo de escritura basado únicamente en la frecuencia de uso de palabras no es confiable, ya que, en cierto sentido para un autor experimentado, es muy simple manipular las palabras de un texto. Por otro lado, las palabras por si solas aportan indicios del tópico del documento. En consecuencia, en lugar de realizar un análisis de estilo se estaría realizando un análisis de los temas sobre los cuales un autor escribe regularmente.

Todo análisis de estilo de escritura parte de la selección de características estilométricas (o marcadores de estilo). Algunos ejemplos de marcadores de estilo son los caracteres del alfabeto, signos de puntación, palabras, etiquetas POS entre otros. Un escritor experto dispone, además de las palabras, de otros recursos del lenguaje que se encuentran en el nivel sintáctico y semántico. En el nivel semántico existen las categorías de las palabras: 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.

A partir de los dos niveles descritos previamente, es posible generar otros tipos de marcadores de estilo, como por ejemplo n-gramas de palabras, n-gramas de etiquetas POS, n-gramas sintácticos. De estos tres tipos n-gramas se generan combinaciones de ellos: palabras-etiquetas POS, etiquetas POS–relaciones de dependencia, palabras-relaciones de dependencia y así sucesivamente. Resulta evidente que diferencia de las palabras, la manipulación consciente de este tipo de marcadores de estilo resulta compleja incluso al autor más experimentado. En consecuencia, estos marcadores permiten análisis de estilo mucho más confiables. Idealmente, se investigan marcadores que identifiquen a un autor sin importar el tipo de documento que este escriba y el tema que aborde.

Para obtener marcadores de estilo robustos, se requieren herramientas computacionales que analizan en lenguaje a nivel de su estructura sintáctica, específicamente a nivel de oraciones. Dicha estructura contiene información muy valiosa sobre la forma en que el autor compone sus oraciones. Eventualmente la estructura (que de forma inconsciente) de las oraciones revela patrones de construcción que no son perceptibles al nivel de los caracteres y las palabras.

Como ejemplos de las herramientas de software para análisis sintáctico se tienen: Stanford Parser, Spacy y Stanza. Sin embargo, es importante recordar que, a pesar de ser herramientas de última generación, incluso entrenadas por medio de aprendizaje profundo, presentan un margen de error que es importante considerar. Particularmente, dichas herramientas presentan detalles en oraciones ambiguas. O bien, que los algoritmos para determinar dónde inicia y donde termina una oración son diferentes entre ellas.

Conclusión.

Se analizaron novelas de 10 autores de habla inglesa; dichas novelas se ordenaron de forma cronológica de acuerdo con la fecha de publicación y se definieron 2 etapas cada una con 3 novelas. Las novelas fueron divididas en oraciones; posteriormente cada novela se dividió en 4 partes proporcionales de acuerdo a la cantidad de oraciones. Una vez generados los textos con distinto número de oraciones, se obtuvieron n-gramas de palabras y etiquetas POS con valores para n= {1, 2, 3} con sus respectivas frecuencias de uso. Los n-gramas y sus frecuencias se almacenaron en matrices de dos dimensiones (Modelo Espacio Vectorial). Una vez que los datos están en las matrices, los n-gramas se ordenaron de forma descendente: de los n-gramas más frecuentemente utilizados a los menos frecuentes.

Dichas matrices se utilizaron para entrenar un algoritmo de clasificación supervisada conocido como Regresión Logística. La estrategia para el entrenamiento fue los *k* marcadores utilizados con mayor frecuencia donde *k*={100, 200, 300, 400, 500}. Adicionalmente se realizaron experimentos con todas las características o marcadores de estilo disponibles en cada configuración.

La etapa del aprendizaje tiene la finalidad de crear un modelo predictivo que sea capaz de determinar a cuál de las dos etapas de un autor pertenece un texto de ejemplo. La métrica para evaluar la eficiencia de los dos tipos de n-gramas y longitudes fue la exactitud (accuracy).

De forma general se encontró que al seleccionar los 100 o 200 n-gramas utilizados con más frecuencia, se obtiene una exactitud que supera a los valores restantes de *k*. Esta tendencia se observó en los n-gramas de palabras y de etiquetas POS así como en los distintos valores de *n*. Utilizar todas las características no presentó ninguna mejora significativa con respecto a los valores de *k.*

Por otro lado, la idea de dividir las novelas en partes es proporcionales tiene dos objetivos: el primero es aumentar de forma artificial el número de textos de cada autor. El segundo es para observar qué ocurre cuando en los textos hay menos cantidad de información: novelas completas (1), medias novelas (2), tercios de novelas (3) y cuartos de novelas (4). Cuando se evalúan novelas completas (1) cada autor cuenta con 6 ejemplos. Mientras que al evaluar cuartos de novelas (4) disponen de 24. Es evidente que a menor cantidad de texto menor es el valor de la métrica exactitud. De acuerdo con los resultados de los experimentos, se sugiere dividir los textos en cuatro bloques de distinto tamaño para observar claramente la tendencia registrada por un autor ante los dos tipos de n-gramas y distintos valores de n.

Finalmente, el cambio de estilo de escritura en algunos autores es más evidente que en otros. En n-gramas de palabras, los autores con la tasa de clasificación más alta fueron Booth Tarkington (BT), Edgar Rice (ER) y John Buchan (JB). La misma tendencia se observó en n-gramas de etiquetas POS. En el otro extremo se encuentran los autores Frederick Marryat (FM) y George Macdonald (GM) particularmente en 1-gramas y 2-gramas de palabras. El mismo caso ocurrió para George Vaisey (GV) pero en n-gramas de etiquetas POS en novelas completas (1) y en medias novelas (2).

Créditos.

Los autores agradecen al Tecnológico Nacional de México por el financiamiento del proyecto 10849.21-P, de la convocatoria de apoyo a proyectos de desarrollo tecnológico e innovación 2021 y las facilidades del Tecnológico Nacional de México campus Tuxtla Gutiérrez para la realización de este trabajo.

Referencias bibliográficas.

Escalante, H., Solorio, T., & Montes, M. (2011). *Local histograms of character n-grams for authorship attribution. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies,* 288-298.

García, V., Mollineda, R., & Sánchez, J. (2009). Index of balanced accuracy: A performance measure for skewed class distributions. *Iberian Conference on Pattern Recognition and Image Analysis* (págs. 441-448). Springer.

Gómez-Adorno, H., Posadas-Durán, J., Ríos-Toledo, G., Sidorov, G., & Sierra, G. (2018). Stylometry-based approach for detecting writing style changes in literary texts. *Computación y Sistemas 22(1),* 47-53.

Houvardas, J., & Stamatatos, E. (2006). N-gram feature selection for authorship Identification. *International Conference on Artifitial Intelligence: Methodology, Systems, and Applications* (págs. 77-86). Springer.

Juola, P. (2008). *Authorship Attribution vol. 3.* Now Publishers Inc.

Sapkota, U., Solorio, T., Montes, M., Bethard, S., & Rosso, P. (2014). Crosstopic authorship attribution: Will out-of-topic data help? *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 1228-1237.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR) 34(1)*, 1-47.

Sidorov, G. (2013). México D.F.: Sociedad Mexicana de Inteligencia Artificial.

Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hernández, L. (2014). Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications 41(3)*, 853-860.

Singhal, A., Salton, G., & Buckley, C. (1995). Length normalization in degraded text collections. *Technical report, Cornell University*.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology 60 (3)*, 538-556.

Toledo, G. R., Sánchez, N., Sidorov, G., & Durán, J. (2019). Identificación de cambios en el estílo de escritura literaria con aprendizaje automático. *Onomázein: Revista de Lingüística, filología y traducción de la Pontificia Universidad Católica de Chile (46)*, 102-128.

Zuo, C., Zhao, Y., & Banerjee, R. (2019). Style change detection with feed-forward neural networks. *CLEF (Working Notes)*.

Información de los autores.

Germán Ríos Toledo obtuvo el grado de Doctor en Ciencias de la Computación en 2019 por el Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET) en Cuernavaca, Morelos, México. Actualmente, es profesor de tiempo completo en el Departamento de Computación del Tecnológico Nacional de México (campus Tuxtla Gutiérrez, Chiapas) en la Ingeniería en Sistemas Computacionales y en la Maestría en Ciencias en Ingeniería Mecatrónica. Su área de especialización es el Procesamiento del Lenguaje Natural, particularmente en el uso de información sintáctica como una característica para el análisis de estilo de escritura. Otras áreas de su interés incluyen el procesamiento y análisis de imágenes, audio y video por medio de Algoritmos de Aprendizaje Automático y Aprendizaje Profundo. Miembro del Sistema Nacional de Investigadores, nivel C (2024-2021).

Cesar Alejandro Meza Pérez, estudiante de Ingeniería en Sistemas Computacionales en el instituto Tecnológico de Tuxtla Gutiérrez, sus áreas de interés se relacionan con las ciencias de la computación, desarrollo de software y desarrollo de aplicaciones móviles.

Jonathan Velázquez Trinidad, estudiante de ingeniería en sistemas computacionales en el Instituto Tecnológico de Tuxtla Gutiérrez, sus áreas de interés se relacionan con base de datos, telecomunicaciones, redes de computadoras y multimedia.

Héctor Guerra Crespo es egresado del I. T. de Mérida (Yucatán, México) de la carrera de Ingeniería en Sistemas Computacionales en 1994, es Doctor en Sistemas Computacionales por la Universidad del Sur (Chiapas, México) en 2011. Es profesor en el área de Ingeniería en Sistemas Computacionales del Instituto Tecnológico de Tuxtla Gutiérrez y en el área de Licenciatura en Sistemas Computacionales de la Universidad Autónoma de Chiapas, en ambas desde 1995. Es miembro del Claustro Doctoral "Doctorado en Ciencias de la Ingeniería" perteneciente al Programa Nacional de Posgrados de Calidad, I.T. de Tuxtla Gutiérrez desde 2016. Miembro del Sistema Nacional de Investigadores, nivel C (2024-2021).

Galdino Belizario Nango Solís, Es Ingeniero en Sistemas Computacionales por el Instituto Tecnológico de Tuxtla Gutiérrez, Chiapas, en 1996. Es Maestro en Ciencias de la Computación por el Centro de Investigación en Computación del Instituto Politécnico Nacional, en 2001. Es Doctor en Desarrollo Tecnológico por la Universidad de Ciencia y Tecnología Descartes, en 2016. Es profesor con 30 Hrs. en el área de Ingeniería en Sistemas Computacionales del Instituto Tecnológico de Tuxtla Gutiérrez, donde imparte clases en las materias de: Programación Lógica y funcional; y Lenguajes y Autómatas. Actualmente es presidente de la Academia de ISC en el ITTG.

Aída Guillermina Cossío Martínez es Maestra en Ciencias en Administración por el Instituto Tecnológico de Tuxtla Gutiérrez en 2002. Es profesora de tiempo completo del área de Ingeniería en Sistemas Computacionales del Instituto Tecnológico de Tuxtla Gutiérrez, desde 1994. Se especializa en la formulación y evaluación de proyectos, así como el emprendimiento y desarrollo de planes de negocio, actualmente es perfil deseable y trabaja en la línea de investigación Tecnología de Información y Base de Datos.