# Propuesta de un análisis estilométrico para las comedias colaboradas de Rojas

Alberto Lara Ramírez
(Universidad de Castilla-La Mancha)

## Introducción

Gracias a las características del paquete *Stylo* de *R*, la investigación del teatro aurisecular cuenta con una nueva herramienta avanzada para estudiar los problemas de autoría. El grupo ETSO ha elaborado informes estilométricos punteros con el propósito de subsanar atribuciones tradicionales erróneas y de desencallar investigaciones en punto muerto. Se trata de hacer justicia con los dramaturgos del siglo XVII. En el presente trabajo, se explora una nueva característica de la librería *Stylo* llamada *General Imposters*, aplicada a las comedias colaboradas de Francisco de Rojas Zorrilla. Así, se pretende proponer un método para desenmascarar la autoría de un texto, utilizando el menor corpus de obras posibles sin dejar de lado la fiabilidad, que siempre debe ir unida a los procedimientos tradicionales.

## 1. Estado de la cuestión

La aplicación *General Imposters* (GI), o *Segundo Sistema de Verificación* (o2), consiste en contrastar un texto dubitado con textos indubitados de posibles candidatos a la autoría y con una selección de textos "impostores" –autores incapaces de haber escrito el texto dubitado–. Kopel y Winter (2014) propusieron el método y dos años después se aplicó para estudiar la autoría de los escritos de Julio César (Kestermont *et al.* 2016). Nosotros ya hemos utilizado la aplicación *General Imposters* en los estudios del teatro del Siglo de Oro, con el fin de certificar la falta de afinidad de *La hermosura y la desdicha* con la producción dramática de Francisco de Rojas.

Por tanto, el objetivo del procedimiento no se basa en determinar la afinidad en estilo de escritura de dos textos en base a la frecuencia de palabras, sino en determinar si dos textos son más afines entre sí que entre otros textos (Eder 2012; Stamatos 2006). En otras palabras, si se pretende demostrar que dos textos X e Y salieron de la pluma del mismo autor, conviene elaborar un corpus de "impostores"; así, el procedimiento determina si X es más afín a Y que al resto de textos "impostores" (Kopel y Winter 2014).

El programa asigna a cada autor candidato una puntuación comprendida entre 0 y 1: 0 corresponde a la exclusión absoluta y 1 a la autoría fiable. Por tanto, teóricamente, una puntuación superior a 0.5 debería resultar suficiente para afirmar con fiabilidad la autoría de un candidato. Sin embargo, hay que tomar con prudencia las puntuaciones comprendidas entre 0.35 y 0.65; se trata de una área gris donde las puntuaciones "parecen sugerir que el clasificador tuvo problemas para tomar decisiones claras" (Eder 2018). Por esa razón, cuando estudiemos el rendimiento de la aplicación en el presente análisis, unos resultados quedarán indeterminados, puesto que la puntuación se encuentra en la mencionada área gris y la prudencia obliga a abstenerse de extraer una conclusión. Existe una función complementaria, "imposters.optimize()", destinada a proponer las puntuaciones que definen esa área gris.

En resolución, el método *General Imposters* se basa en la similitud, ya que el funcionamiento consiste en medir la afinidad entre el texto indubitado y el resto de textos: el texto indubitado se atribuye a un candidato a la autoría si la escritura es similar.

## 2. Establecimiento del corpus

El único candidato a la autoría con que se trabajará es Francisco de Rojas Zorrilla para determinar si los textos evaluados pertenecen o no a su pluma. Por ese motivo, para

que el programa pueda establecer la relación, se necesita un corpus de comedias indubitadas de Rojas. Se utilizarán 27 comedias, donde la crítica considera fiable la autoría de Rojas. Las comedias son *Abrir el ojo*; *Cada cual lo que le toca*; *Casarse por vengarse*; *Donde hay agravios no hay celos*; *El Caín de Cataluña*; *El más impropio verdugo*; *El profeta falso Mahoma*; *Entre bobos anda el juego*; *La traición busca el castigo*; *Lo que quería ver el marqués*; *Lo que son mujeres*; *Los áspides de Cleopatra*; *Los bandos de Verona*; *Los celos de Rodamonte*; *Los encantos de Medea*; *Los trabajos de Tobías*; *Morir pensando matar*; *No hay amigo para amigo*; *No hay ser padre siendo rey*; *Nuestra señora de Atocha*; *Obligados y ofendidos*; *Peligrar en los remedios*; *Persiles y Segismunda*; *Primero es la honra que el gusto*; *Progne y Filomena*; *Santa Isabel reina de Portugal*; y *Sin honra no hay amistad*.

Para el análisis del texto dubitado, el programa considera colectivamente las 27 obras como un único documento. Así, se consigue una evaluación eficaz del texto dubitado, estableciendo si se relaciona con la escritura del dramaturgo toledano más que con el resto de textos. Por tanto, ahora el desafío del proceso radica en la adecuada selección de los "impostores". De entrada, si se pretende conseguir una representatividad estadística fiable, los textos del corpus deben constar de un número de palabras suficientes. Se ha demostrado que el método proporciona resultados fiables para desenmascarar a un autor con textos de 500 palabras (Koppel y Winter 2014: 179). Por lo general, las obras de teatro auriseculares, con alrededor de 3000 versos, no supone un obstáculo para la aplicación del método. Incluso, también las jornadas de las comedias, por separado, cumplen el requisito; se trata de un dato importante porque nosotros trabajaremos con jornadas, en lugar de comedias completas, puesto que el objetivo consiste en rastrear la pluma de Rojas en comedias con jornadas de varios ingenios.

Como el único candidato por evaluar es Rojas, el resto de comedias se unificarán en una categoría llamada "Otro". Así, si el programa no encuentra afinidad con Rojas, determinará que la comedia pertenece a "Otro". En dicha categoría, se encuentran autores que han podido escribir la obra dubitada y autores que no la han podido escribir –los "impostores"–. Respecto al número de "impostores", parafraseando a Koppel y Winter (2014, 186), se requiere un equilibrio adecuado entre la cantidad y la calidad. Cuando se habla de la calidad de los "impostores", se refiere a la similitud con el texto dubitado: se aconseja que cuenten con un número similar de palabras y pertenezcan al mismo género. Como los "impostores" serán también comedias auriseculares, contarán con una elevada cantidad, por lo que no será necesario incluir un elevado número. Así, se cumple una de las principales pretensiones del análisis: crear un método estilométrico preciso con el menor corpus posible.

Con el objetivo de que la categoría "Otro" tenga una cantidad de obras iguales que la categoría "Rojas", se han seleccionado 27 obras de diferentes autores. Así, el programa tendrá una amplia visión de la escritura de múltiples dramaturgos auriseculares y podrá separar adecuadamente el estilo de Rojas. Las 27 obras son las siguientes: *Adonis y Venus*; *Casa de dos puertas mala es de guardar*; *Del rey abajo, ninguno*; *Deste agua no beberé*; *El anzuelo de Fenisa*; *El caballero de Olmedo*; *El conde de Sex*; *El condenado por desconfiado*; *El desdén con el desdén*; *El esclavo del demonio*; *El lindo don Diego*; *El mágico prodigioso*; *El médico de su honra*; *El valiente justiciero*; *El vergonzoso en palacio*; *La dama boba*; *La dama duende*; *La mayor corona*; *La serrana de la Vera*; *La verdad sospechosa*; *La vida en el ataúd*; *Las bizarrías de Belisa*; *Los cabellos de Absalón*; *Marta la piadosa*; *No puede ser guardar una mujer*; *Reinar después de morir*; y *Tan largo me lo fiais*.

## 3. Pruebas de control

Antes de iniciar el análisis propuesto, conviene realizar unas pruebas de control para comprobar que el corpus se ha seleccionado correctamente y el método funciona. Por una parte, la primera tabla ofrece los resultados de analizar las 27 comedias indubitadas de Rojas por jornadas. Se ha evaluado cada jornada de las 27 comedias, comparándola con el resto de jornadas de las comedias indubitadas de Rojas y con las piezas incluidas en la categoría "Otro". El informe, salvo en tres jornadas, determina acertadamente que las obras pertenecen a Rojas, como establece el consenso de la crítica; asimismo, en las tres jornadas conflictivas no se descarta a Rojas, sino que se mantiene en la interminación.

| **Nombre** | **Primera jornada** | | **Segunda jornada** | | **Tercera jornada** | |
|---|---|---|---|---|---|---|
| | **Rojas** | **Otro** | **Rojas** | **Otro** | **Rojas** | **Otro** |
| *Abrir el ojo* | 1 | 0.16 | 1 | 0.09 | 1 | 0.11 |
| *Cada cual lo que le toca* | 1 | 0.01 | 1 | 0.04 | 1 | 0 |
| *Casarse por vengarse* | 1 | 0.07 | 0.96 | 0.16 | 1 | 0 |
| *Donde hay agravios* | 1 | 0.08 | 1 | 0.02 | 1 | 0 |
| *El Caín de Cataluña* | 1 | 0.11 | 1 | 0.02 | 1 | 0.10 |
| *El más impropio verdugo* | 0.74 | 0.65 | 0.91 | 0.33 | 0.98 | 0.10 |
| *El profeta falso Mahoma* | 1 | 0.03 | 0.98 | 0.11 | 1 | 0.19 |
| *Entre bobos anda el juego* | 0.96 | 0.31 | 1 | 0.07 | 1 | 0.20 |
| *La traición busca el castigo* | 1 | 0.09 | 1 | 0.01 | 1 | 0.02 |
| *Lo que quería ver el marqués* | 1 | 0.19 | 1 | 0.27 | 1 | 0.12 |
| *Lo que son mujeres* | 0.97 | 0.28 | 0.96 | 0.24 | 1 | 0.22 |
| *Los áspides de Cleopatra* | 1 | 0.08 | 1 | 0 | 1 | 0.03 |
| *Los bandos de Verona* | 1 | 0.03 | 1 | 0.01 | 1 | 0.01 |
| *Los celos de Rodamonte* | 1 | 0.01 | 1 | 0.02 | 1 | 0.02 |
| *Los encantos de Medea* | 1 | 0.09 | 0.98 | 0.25 | 0.96 | 0.11 |
| *Los trabajos de Tobías* | 1 | 0.19 | 1 | 0.17 | 0.97 | 0.29 |
| *Morir pensando matar* | 0.87 | 0.50 | 0.90 | 0.35 | 0.82 | 0.55 |
| *No hay amigo para amigo* | 1 | 0.09 | 1 | 0.01 | 1 | 0 |
| *No hay ser padre siendo rey* | 1 | 0.03 | 1 | 0.04 | 1 | 0.03 |
| *Nuestra señora de Atocha* | 1 | 0.09 | 1 | 0.02 | 1 | 0.14 |
| *Obligados y ofendidos* | 1 | 0.14 | 1 | 0.01 | 1 | 0.22 |
| *Peligrar en los remedios* | 1 | 0.01 | 1 | 0.02 | 1 | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| *Persiles y Segismunda* | 1 | 0.01 | 1 | 0.06 | 1 | 0.04 |
| *Primero es la honra que el gusto* | 0.96 | 0.4 | 1 | 0.34 | 1 | 0.17 |
| *Progne y Filomena* | 1 | 0 | 1 | 0 | 1 | 0.04 |
| *Santa Isabel reina de Portugal* | 0.98 | 0.19 | 0.97 | 0.17 | 1 | 0 |
| *Sin honra no hay amistad* | 1 | 0.01 | 1 | 0.02 | 1 | 0.03 |

Por otra parte, la segunda tabla ofrece los resultados de analizar 27 comedias de otros autores por jornadas. De nuevo, el informe determina acertadamente que las jornadas no pertenecen a Rojas, salvo en tres ocasiones, donde el programa no puede determinar nada. Por tanto, sumando estos resultados a los de la primera tabla, se supera la prueba, ya que se ha comprobado que el programa, con el corpus seleccionado, es capaz de distinguir con claridad el estilo de Rojas.

| **Nombre** | **Primera jornada** | | **Segunda jornada** | | **Tercera jornada** | |
|---|---|---|---|---|---|---|
| | **Rojas** | **Otro** | **Rojas** | **Otro** | **Rojas** | **Otro** |
| *Adonis y Venus* | 0.51 | 0.90 | 0.30 | 0.96 | 0.32 | 0.97 |
| *Casa de dos puertas* | 0.19 | 1 | 0.07 | 1 | 0.05 | 1 |
| *Del rey abajo, ninguno* | 0.16 | 1 | 0.15 | 1 | 0.37 | 1 |
| *Deste agua no beberé* | 0.05 | 1 | 0.02 | 1 | 0.11 | 1 |
| *El anzuelo de Fenisa* | 0 | 1 | 0.01 | 1 | 0 | 1 |
| *El caballero de Olmedo* | 0 | 1 | 0 | 1 | 0 | 1 |
| *El conde de Sex* | 0.96 (0.47) | 0.44 (0.38)[1] | 0.26 | 0.97 | 0.31 | 0.97 |
| *El condenado por desconfiado* | 0.10 | 1 | 0.14 | 1 | 0.10 | 1 |
| *El desdén con el desdén* | 0.13 | 1 | 0 | 1 | 0.08 | 1 |
| *El esclavo del demonio* | 0.07 | 1 | 0.08 | 1 | 0.04 | 1 |
| *El lindo don Diego* | 0.09 | 1 | 0.28 | 1 | 0.13 | 1 |
| *El mágico prodigioso* | 0.08 | 1 | 0.20 | 1 | 0.06 | 1 |
| *El médico de su honra* | 0.24 | 0.98 | 0.35 | 0.94 | 0.18 | 1 |
| *El valiente justiciero* | 0.19 | 1 | 0.20 | 1 | 0.15 | 1 |

[1] La puntuación entre paréntesis responde al resultado de delimitar el área gris con la función "imposters.optimize()". Así pues, con *El conde de Sex*, por un lado, se considera fiable un resultado superior a 0.47, que se obtiene gracias a ese 0.97; sin embargo, por otro lado, se considera descartable una autoría con un resultado inferior a 0.38, que no se obtiene debido a ese 0.44. Por tanto, al no descartar por completo la opción "Otro", no se puede considerar la autoría de Rojas, de modo que la prudencia obliga a mantener la indeterminación.

| | | | | | | |
|---|---|---|---|---|---|---|
| *El vergonzoso en palacio* | 0.09 | 1 | 0.06 | 1 | 0.02 | 1 |
| *La dama boba* | 0.05 | 1 | 0.04 | 1 | 0 | 1 |
| *La dama duende* | 0.06 | 1 | 0.10 | 1 | 0.09 | 1 |
| *La mayor corona* | 0.01 | 1 | 0.08 | 1 | 0.08 | 1 |
| *La serrana de la Vera* | 0.03 | 1 | 0.07 | 1 | 0.05 | 1 |
| *La vida en el ataúd* | 0.11 | 1 | 0.06 | 0.97 | 0.06 | 1 |
| *La verdad sospechosa* | 0.15 | 1 | 0.06 | 1 | 0.13 | 1 |
| *Las bizarrías de Belisa* | 0.03 | 1 | 0.08 | 1 | 0.02 | 1 |
| *Los cabellos de Absalón* | 0.26 | 0.97 | 0.23 | 0.92 | 0.26 | 0.97 |
| *Marta la piadosa* | 0.02 | 1 | 0.23 | 1 | 0.02 | 1 |
| *No puede ser guardar una mujer* | 0.06 | 1 | 0.10 | 1 | 0.15 | 1 |
| *Reinar después de morir* | 0.39 | 0.93 | 0.23 | 1 | 0.19 | 1 |
| *Largo me lo fiais* | 0 | 1 | 0 | 1 | 0 | 1 |

## 4. Análisis de las comedias colaboradas

A continuación, se ofrece una tabla con el resultado de utilizar el procedimiento con las comedias colaboradas donde la crítica ha señalado la intervención de Rojas. El propósito es comprobar si el análisis consigue descubrir el estilo del dramaturgo toledano. Después del cuadro sinóptico, se interpreta individualmente cada obra.

| **Nombre** | **Primera jornada** | | **Segunda jornada** | | **Tercera jornada** | |
|---|---|---|---|---|---|---|
| | **Rojas** | **Otro** | **Rojas** | **Otro** | **Rojas** | **Otro** |
| *El catalán Serrallonga* | 0.61 | 0.70 | 1 | 0.08 | 0.63 | 0.82 |
| *El jardín de Falerina* | 1 | 0.04 | 0.69 | 0.61 | 0.40 | 0.87 |
| *El mejor amigo, el muerto* | 0.67 | 0.76 | 0.98 | 0.05 | 0.56 | 0.92 |
| *El monstruo de la fortuna* | 0.38 | 0.90 | 0.48 | 0.79 | 1 | 0.06 |
| *El pleito del diablo* | 0.35 | 0.95 | 0.14 | 0.99 | 0.49 | 0.90 |
| *El robo de las sabinas* | 0.85 (0.42) | 0.44 (0.31) | 1 | 0.03 | 0.69 (0.34) | 0.78 (0.53) |
| *El villano gran señor y gran Tamorlán de Persia* | 0.77 | 0.61 | 0.83 | 0.64 | 0.95 | 0.35 |
| *Empezar a ser amigos* | 0.81 | 0.51 | 0.98 | 0.12 | 0.86 | 0.46 |
| *La Baltasara* | 0.20 | 0.93 | 0.77 | 0.72 | 0.59 | 0.80 |
| *La fingida Arcadia* | 0.77 | 0.56 | 0.98 (0.65) | 0.42 (0.23) | 0.20 | 0.98 |

| | | | | | | |
|---|---|---|---|---|---|---|
| *La más hidalga hermosura* | 0.91 (0.46) | 0.37 (0.44) | 0.77 | 0.55 | 0.35 | 0.88 |
| *Los privilegios de las mujeres* | 0.99 | 0.21 | 1 | 0.17 | 0.56 | 0.88 |
| *También la afrenta es veneno* | 0.70 | 0.75 | 0.87 | 0.45 | 1 | 0 |
| *También tiene el sol menguante* | 0.60 | 0.87 | 0.91 | 0.31 | 0.78 | 0.63 |

1. *El catalán Serrallonga*: el informe respalda la autoría de Rojas de la segunda jornada, pero es incapaz de determinar la primera, atribuida a Antonio Coello, y la tercera jornada, atribuida a Vélez de Guevara.
2. *El jardín de Falerina*: el informe respalda la autoría de Rojas de la primera jornada, y la autoría de otro autor de la tercera jornada, atribuida a Calderón, pero es incapaz de determinar la segunda, atribuida a Antonio Coello.
3. *El mejor amigo, el muerto*: el informe respalda la autoría de Rojas de la segunda jornada, pero quedan indeterminadas la primera, atribuida a Belmonte Bermúdez, y la tercera jornada, atribuida a Calderón.
4. *El monstruo de la fortuna*: el informe respalda la autoría de Rojas de la tercera jornada, y la autoría de otro autor de la primera jornada, atribuida a Calderón, pero es incapaz de determinar la segunda, atribuida a Antonio Coello.
5. *El pleito del diablo*: el informe respalda que la primera jornada, atribuida a Vélez, no pertenece a Rojas. En cuanto a la segunda jornada, atribuida al dramaturgo toledano, el informe determina que pertenece a otro autor con claridad. La tercera jornada, atribuida a Mira, queda indeterminada.
6. *El robo de las sabinas*: el informe respalda la autoría de Rojas de la segunda jornada, pero es incapaz de determinar la primera, atribuida a Antonio Coello, y la tercera jornada, atribuida a Juan Coello.
7. *El villano gran señor y gran Tamorlán de Persia*: el informe determina la autoría de Rojas de la tercera jornada, atribuida a Roa, y es incapaz de determinar la primera, atribuida a Rojas, y la segunda jornada, atribuida a Villanueva.
8. *Empezar a ser amigos*: el informe respalda la autoría de Rojas de la segunda jornada. La primera y la tercera quedan indeterminadas.
9. *La Baltasara*: el informe respalda que la primera jornada, atribuida a Vélez, no pertenece a Rojas. La segunda jornada, atribuida a Antonio Coello, queda indeterminada, como la tercera jornada, atribuida a Rojas.
10. *La fingida Arcadia*: el informe respalda la autoría de otro autor de la tercera jornada. Quedan indeterminadas la primera y la segunda.
11. *La más hidalga hermosura*: quedan indeterminadas las tres jornadas, atribuidas a Zabaleta, Rojas y Calderón, respectivamente.
12. *Los privilegios de las mujeres*: el informe respalda la autoría de Rojas de la primera jornada, atribuida a Antonio Coello, y la segunda jornada. La tercera, atribuida a Calderón, queda indeterminada.
13. *También la afrenta es veneno*: el informe respalda la autoría de Rojas de la tercera jornada, pero es incapaz de determinar la primera y la segunda jornada, atribuidas a Vélez y a Antonio Coello, respectivamente.
14. *También tiene el sol menguante*: el informe respalda la autoría de Rojas de la segunda jornada, pero es incapaz de determinar la primera y la tercera jornada.

## 5. Conclusión

En la mayoría de las ocasiones, el método identifica sin problemas a Rojas en las jornadas donde la crítica ha supuesto la intervención de su pluma. Por tanto, se reafirma la conclusión establecida tras las pruebas de control: el análisis reconoce con claridad a Rojas. Sin embargo, destaca el elevado número de indeterminación en las comedias colaboradas respecto a las comedias de un autor. El método identifica con facilidad el estilo de Calderón de la Barca, además del estilo del dramaturgo toledano. Sin embargo, no reconoce apenas a Vélez de Guevara y nunca reconoce a los hermanos Coello, Mira de Amescua y Zabaleta.

**Obras citadas**

Cuéllar, Álvaro y Vega García-Luengos, Germán. ETSO: Estilometría aplicada al Teatro del Siglo de Oro. 2017-2024. Recurso web <http://etso.es/>.https://doi.org/10.17613/a2f6-1y65.

Eder, M. "Autorship verification with the package 'stylo'." *Computational Stylistics Group*. Recurso web: https://computationalstylistics.github.io/blog/imposters/.

---. "Computational stylistics and Biblical translation: How reliable can a dendogram be?" Piotrowski, T. y Grabowski, Ł. eds. *The Translator and the Computer*. Wrocław: WSF Press, 2012. 155–70

Eder, M., Rybicki, J. y Kestemont, M. "Stylometry with R. A package for computational text analysis." *R Journal* 8(1) (2016): 107-121.

Kestermont M., Stover, J., Koppel, M., Karsdorp, F. y Daelemans, W. "Authenticating the writings of Julius Caesar." *Expert Systems with Applications* 63 (2016): 86-96.

Koppel, M., y Winter, Y. "Determining if two documents are written by the same author." *Journal of the Association for Information Science and Technology* 65(1) (2014): 178-187.

Sanderson, C. y Guenter, S. "Short text authorship attribution via sequence kernels, Markov chains and author unmasking: an investigation." *Proceedings of International Conference on Empirical Methods in Natural Language Processing*. EMNLP, 2006. 482-491.

Stamatatos, E. "Authorship attribution base don feature set subspacing ensembles." *International Journal on Artificial Intelligence Tools* 15(05) (2006): 823-838. DOI: 10.1142/S0218213006002965.