Revista de Humanidades Digitales

Clasificación de tragedias y comedias en las comedias nuevas de Calderón de la Barca[1]

Classification of Tragedies and Comedies in Calderón de la Barca's Comedias Nuevas

Jörg LEHMANN
Universidad Eberhard Karls de Tubinga
joerg.lehmann@romanistik.uni-tuebingen.de
https://orcid.org/0000-0003-1334-9693

Sebastian PADÓ
Universidad de Stuttgart
sebastian.pado@ims.uni-stuttgart.de
https://orcid.org/0000-0002-7529-6825

RESUMEN

El objetivo de este estudio es clasificar 112 dramas escritos por Calderón de la Barca, comedias y tragedias, utilizando procedimientos computacionales basados en la semántica distribucional. Quince de estas *comedias nuevas* ya han sido clasificadas cualitativamente por investigadores especialistas como tragedias o comedias; para otros 82 dramas no había datos sobre su clasificación. En este artículo exploramos cuatro métodos independientes de *document embedding* que difieren entre sí, por un lado, en la creación y reducción de la matriz de rasgos y, por otro lado, en el cálculo de las matrices de similitud o distancia. Los mejores resultados –medidos con respecto a los dramas clasificados manualmente– se obtienen mediante el procedimiento de clasificación que aplica la reducción de información más compleja en la matriz de rasgos. Además, se lleva a cabo un análisis contrastivo de vocabulario con *word embeddings*. Aquí se comparan dos subcorpus que contienen obras de teatro clasificadas de manera manual y se utilizan tanto las listas de palabras producidas por los cuatro métodos probados o mediante la distribución de probabilidad *log-likelihood*. Este paso permite identificar 130 términos que distinguen entre comedias y tragedias. El resultado muestra que los métodos explorados identifican las tragedias con mayor precisión que las comedias, lo que indica que las primeras tienen más rasgos distintivos. También se hace evidente que se podrían considerar más adecuadamente clasificaciones como tragedia y comedia como polos de un espectro entre los que se pueden observar diferencias graduales, por lo que la zona de transición resultante contiene comedias nuevas (que han sido descritas en investigaciones anteriores como tragicomedias) o comedias mitológicas.

PALABRAS CLAVE

Análisis de clúster, análisis de dramas, clasificación, Pedro Calderón de la Barca, Siglo de Oro.

ABSTRACT

In this study, we aim at distinguishing comedies and tragedies among 112 dramas written by Calderón de la Barca, using procedures established by distributional semantics. Fifteen of these *comedias nuevas* have already been classified by qualitative researchers as either tragedies or comedies, respectively; for another 82 dramas the classification was unknown. Four independent document embedding methods are explored, which differ from each other in matrix creation and reduction, and in the calculation of similarity or distance matrices. The best results –measured against the pre-established classification of these dramas– are obtained through the classification procedure that applied the strongest matrix reduction. In addition, a contrastive vocabulary analysis with word embeddings is carried out, based either on word lists produced by the four tested methods, or on the log-likelihood probability distribution for two sub-corpora containing only dramas already determined to be comedies or tragedies. This step permits the identification of 130 terms that are each discriminative either of comedies or of tragedies. The outcome shows that the explored methods identify tragedies with greater accuracy than comedies, indicating that tragedies have more distinctive features. It also becomes apparent that one could more appropriately consider classifications such as tragedy and comedy as poles between which gradual differences can be observed, whereby the ensuing transitional area contains *comedias nuevas* that have been described in prior research as *tragicomedias* or *comedias mitológicas*.

KEYWORDS

Cluster Analysis, Classification, Drama Analysis, Pedro Calderón de la Barca, Spanish Golden Age.

RHD 7 (2022)

ISSN
2531-1786

10.5944/rhd.vol.7.2022

Recibido 20/09/2022 – Aceptado 19/10/2022

1. INTRODUCCIÓN

Pedro Calderón de la Barca (1600-1681) es, junto con Félix Lope de Vega Carpio (1562-1635), uno de los dramaturgos más importantes del Barroco español, período incluido dentro del conocido como el *Siglo de Oro*. Las obras de este autor incluyen 84 autos sacramentales, 112 comedias y 41 piezas cortas (teatro cómico breve). A principios del siglo xx sus obras teatrales fueron publicadas de manera casi completa por la editorial madrileña Aguilar (Calderón de la Barca, 1951-1956)[2]. Las comedias que se publicaron en vida del autor especificaban los títulos de las obras dramáticas con términos como *gran comedia* o *comedia famosa*. Sin embargo, estas descripciones no diferenciaban entre comedias y tragedias. El término comedia era intercambiable con el de obra o pieza teatral:

> Aunque la etimología de comedia es bastante simple –una obra de teatro de altos espíritus y risas con un final feliz–, en la España de principios de la Edad Moderna el término comedia significaba 'una obra de teatro' u 'obra para la escena' en un sentido bastante neutro (Sullivan, 2018, p. 33).

De esta manera, la mención de *comedia* en el título no diferenciaba entre lo que hoy en día entendemos como comedias y tragedias.

Dado que Calderón nunca escribió una poética, se considera como referencia contemporánea la obra programática de Lope de Vega (1621) *Arte nuevo de hacer comedias en este tiempo* de 1609, según cuyas reglas pragmáticas se orientó Calderón de manera general, a pesar de algunas ligeras modificaciones. En esta poética Lope define la comedia nueva como una obra de teatro en tres actos. El autor constata la comedia como un drama de ficción que involucra a la gente común, lo que la distingue de la tragedia, en la que aparecerían miembros de la familia real o de la nobleza y que estaría basada en hechos históricos. Además, Lope caracteriza la comedia nueva como una mezcla de elementos cómicos y trágicos, refiriéndose así a la combinación de ambos géneros dramáticos[3]. Así pues, los dramaturgos españoles del siglo XVII disponían de una referencia poetológica central que, superando la poética aristotélica, definía un nuevo *estilo español* que se aplicaba tanto a la comedia como a la tragedia. En cuanto a la recepción histórica de la comedia nueva española, esta pasa en primer lugar por una fase de desprestigio, por su irregularidad según las doctrinas del clasicismo francés. Sin embargo, la comedia nueva, y especialmente su forma de entender las tragedias, proporciona posteriormente una influencia vital a través de las etapas históricas alemanas de la Ilustración, el Romanticismo y el Idealismo.

Gotthold Ephraim Lessing (1729-1781) fue uno de los primeros en la región germanoparlante en reconocer el valor de la obra de Calderón. Este autor analizó intensamente las tragedias

[1] Los autores están muy agradecidos al Dr. José Calvo Tello, de la Biblioteca Estatal y Universitaria en Göttingen, Alemania, por la traducción al idioma español. Este estudio surgió como parte del proyecto QUOTE: Modelización integral del habla en los textos en prosa, patrocinado por la Comunidad Alemana de Investigación (Deutsche Forschungsgemeinschaft, proyecto número 350397899). Los autores agradecen al Prof. Dr. Hanno Ehrlicher (Universidad de Tubinga), que comentó la primera versión del artículo.

[2] Esta publicación, sin embargo, no se ajusta a las normas de una edición histórico-crítica.

[3] Esto puede considerarse como una referencia a un tercer género, que ha recibido poca atención hasta ahora en la investigación. Compárese aquí Couderc (2012, pp. 65-75 y 102-109).

del Siglo de Oro español y puso en práctica sus aspiraciones teóricas en un género recién fundado en la literatura alemana: la tragedia burguesa. Otros investigadores románticos siguieron sus estudios sobre Calderón, como Ludwig Tieck, August Wilhelm y Friedrich Schlegel, los hermanos Grimm y Alexander y Wilhelm von Humboldt, que habían todos estudiado español en Gotinga (Sullivan, 2017). August Wilhelm Schlegel tradujo cinco obras de Calderón en su edición de *Teatro Español*, que se publicó en 1803 (el primer volumen) y en 1809 (el segundo) y analizó los textos calderonianos con gran detalle en sus cursos llamados *Vorlesungen über dramatische Kunst und Literatur* –*Conferencias sobre arte dramático y literatura*– en la Universidad de Viena (1809). Wilhelm Joseph Schelling desarrolló su propia teoría de las tragedias en su lección *Abhandlung über die Tragödie* –*Ensayo sobre la tragedia*– basada en la obra de Calderón. Incluso Hegel o Schopenhauer se ocuparon de las obras de Calderón, por lo que no es de extrañar que Walter Benjamin vuelva una y otra vez a Calderón y a su noción de la tragedia en su *Ursprung des deutschen Trauerspiels* –*Origen de la tragedia alemana*– (Benjamin, 1978).

El interés en la región germanoparlante se centraba principalmente en las tragedias de Calderón y, por tanto, en unas pocas obras; fue solo a mediados del siglo XX cuando se hicieron intentos serios de examinar y clasificar el conjunto de las comedias nuevas calderonianas. Fueron inicialmente los editores de las *Obras completas de Calderón* (Valbuena Briones & Astrana Marín, 1951) quienes, en 1951, emprendieron una división binaria de estas piezas teatrales en *dramas* o *serias* (las que se asemejan a las tragedias) y *comedias* o *ligeras* (los dramas orientados al entretenimiento). De este modo, los editores de la editorial Aguilar organizan las comedias de Calderón siguiendo una clara separación entre comedia y tragedia siguiendo la tradición poética desde Aristóteles, pero aplicando para ello criterios poco explícitos[4]. Al mismo tiempo, estos editores crearon una diferenciación que ha sido discutida acaloradamente con posiciones enfrentadas en la investigación literaria de la obra calderoniana desde la segunda mitad del siglo XX hasta la actualidad.

La escuela británica (Alexander A. Parker, Bruce Wardropper, Anthony Irving Watson o Henry W. Sullivan entre otros) se ocupó intensamente de las tragedias calderonianas. Sus intentos de clasificación fueron objeto de una crítica rigurosamente metódica a principios de este milenio por parte del investigador español Jesús G. Maestro (2003), quien comenta, no sin sarcasmo, la "impotencia de la teoría literaria" respecto a los géneros dramáticos y las atribuciones siempre cambiantes que los acompañan[5]. Posteriormente el investigador británico Henry W. Sullivan identifica, desde una perspectiva cualitativa, doce criterios según los cuales se puede caracterizar el drama trágico del Siglo de Oro. Para ello, Sullivan se centra principalmente en los rasgos temáticos –tales como conflictos paternofiliales, dramas basados en la venganza o el honor–, en los indicios extraliterarios –personas de alta posición social–, en las características de la trama –juicios injustos o muerte del protagonista–, en los atributos de la recepción –creación de *eleos* y *pathos* o

[4] Compárese aquí la introducción de Ángel Valbuena Briones en Calderón de la Barca (1951, p. 9-34).
[5] Ver Maestro (2003) y también la discusión de Arellano (2018) sobre los límites de la compilación de taxonomías.

finales catárticos–, o en la formulación de criterios de exclusión –por ejemplo, temas como la redención y la condenación o los dramas de mártires– (Sullivan, 2018, pp. 362-364). En el marco de estos criterios, Sullivan pudo identificar al menos catorce tragedias en las obras completas de las comedias nuevas calderonianas.

A la luz de la monumental obra de Calderón, no es de extrañar que, por un lado, la clasificación de las comedias nuevas nunca se haya realizado de forma exhaustiva quitando la de los editores de la edición de Aguilar[6]: ¿qué investigadora o investigador está dispuesto a estudiar y clasificar 112 dramas? Al mismo tiempo, es evidente que justo este tipo de obras escritas son aptas para la aplicación de procedimientos computacionales. Por otro lado, hay que tener en cuenta que una clasificación computacional basada en todo el corpus de las comedias ha sido imposible hasta la primavera de 2022, cuando todas ellas se publicaron codificadas en formato electrónico[7]. Salvo algunos estudios, las obras de Calderón no han sido todavía abordadas con los métodos que proporcionan las humanidades digitales; sin embargo, es evidente que se prestan al examen de las similitudes estructurales entre las obras de un determinado género o a las diferencias entre los dramas de distintos géneros (Peña-Pimentel, 2011, 2012; de la Rosa et al., 2018; Ehrlicher et al., 2020). La obra de Calderón representa un caso llamativo en el que un autor escribió una cantidad enorme de piezas teatrales en un periodo relativamente corto del siglo XVII.

El estudio que nos ocupa representa un intento de evaluar críticamente la validez de la distinción entre la comedia y la tragedia analizando 112 dramas. Esto se realiza en paralelo a una evaluación de las posibilidades metódicas que ofrece la aplicación de los procedimientos inspirados en la semántica distribucional en las humanidades digitales para este problema[8]. Dado que, hasta ahora, solo se ha estudiado una pequeña parte de las comedias calderonianas, y que la mayoría de ellas permanecen totalmente inexploradas, se espera que los métodos probados puedan aportar indicaciones importantes para la clasificación de las obras que aún no han sido analizadas en profundidad.

2. METODOLOGÍA

2.1. Base metodológica

En la actualidad, el concepto de semántica distribucional se utiliza ampliamente en el ámbito de la lingüística computacional. Esta se basa en que el significado de una forma se establece en

6 Un intento de ello es el portal Calderón Digital (http://calderondigital.tespasiglodeoro.it/), mediante el cual se pueden filtrar alrededor de 80 textos escritos de Calderón según sus características de género; también se incluyen los investigadores responsables de estas clasificaciones.

7 La colección completa está disponible en XML-TEI en DraCor https://dracor.org/cal. No solo se han puesto a disposición las 110 comedias nuevas que figuran en la edición de Aguilar, sino también otras dos comedias atribuidas a Calderón, a saber, La selva confusa y Cómo se comunican dos estrellas contrarias. Para la discusión de esta atribución, véase Coenen (2016). Los autores de este estudio están muy agradecidos al Dr. Simon Kroll y su equipo de la Universidad de Viena por la contribución de más de 50 dramas a este corpus.

8 Estudios comparables sobre el drama clásico francés han sido presentados hasta ahora por, por ejemplo, Christof Schöch (2013, 2017), que abordaron el tema con métodos de modelización tópica y estilométrica. Compárese aquí la introducción de Ángel Valbuena Briones en Calderón de la Barca (1951, pp. 9-34).

función de su uso y de su coaparición con otras formas dentro de un contexto específico; estas coapariciones pueden ser cuantificadas y las palabras y los documentos se representan en un espacio de alta dimensión; las relaciones semánticas se infieren a partir de las similitudes en ese espacio. Para la representación de los documentos, la frecuencia (absoluta o relativa) de las palabras de cada documento se almacena en forma de matrices de vectores en las que cada palabra corresponde a una columna de la matriz y cada documento a una fila. Las celdas de la matriz contienen las frecuencias de coocurrencia. Estas frecuencias totales se suelen transformar para contrarrestar la distribución Zipf[9] de las palabras, por ejemplo, mediante *información mutua puntual* (en inglés, *pointwise mutual information*) o la *tf-idf* (*frecuencia de términos-frecuencia de documentos inversa*) (Lowe, 2001). Para representar los significados de las palabras, se crea el mismo tipo de matriz, con los términos formando filas y las palabras contextuales formando columnas. Estas matrices pueden servir para calcular las distancias entre palabras individuales o textos, para compararlas entre sí, para agruparlas o para visualizarlas. Por regla general, estas matrices contienen miles de columnas y son dispersas o huecas (en inglés, *sparse*, es decir, que muchos de sus valores son ceros). Por ello, es conveniente reducirlas a un número mucho menor de dimensiones para su procesamiento mediante matrices de distancia o similitud. Esta reducción dimensional es un requisito puramente técnico que altera poco los datos subyacentes (Jockers, 2013, pp. 63-67). Los vectores de baja dimensión resultantes suelen denominarse *word embeddings* o *document embeddings* y son probablemente la práctica más habitual como representación aproximadamente semántica en el Procesamiento del Lenguaje Natural (PLN). Este método está relacionado con *topic modeling* aunque no es idéntico.

La elección de un enfoque distributivo para este trabajo se basa en la hipótesis de que las comedias y las tragedias pueden diferenciarse observando selección y uso de las palabras. En pocos términos, cabe esperar que en las tragedias calderonianas coocurran términos como *honor*, *poder* o *muerte*, mientras que en las comedias se tiende a combinar palabras como *amor*, *disfraz* o *celos*. Evidentemente, se trata de un planteamiento que representa una simplificación excesiva: los patrones narrativos o las estructuras argumentales no pueden caracterizarse de esta manera. Al mismo tiempo, el gran éxito de los enfoques basados en frecuencia o coaparición de las palabras, como los métodos habituales de reconocimiento de autores de la estilometría, demuestra que estos enfoques permiten una comprensión sorprendentemente efectiva también de textos literarios.

2.2. Corpus

De las 112 obras dramáticas de Calderón que se tenía a disposición, se creó un subcorpus con obras claramente anotadas como comedias o tragedias. Este subcorpus contiene, por un lado, las catorce tragedias identificadas por Sullivan, a las que se añadió otra más que aparentemente había permanecido desconocida para él: *Saber del bien y del mal* (Escudero Baztán, 2021, p. 21).

[9] La ley de Zipf establece que, dado un corpus de expresiones del lenguaje natural, la frecuencia de cualquier palabra es inversamente proporcional a su rango en la tabla de frecuencias. Véase https://es.wikipedia.org/wiki/Ley_de_Zipf.

Por otro lado, quince dramas identificados por la investigación tradicional como comedias constituyen la contrapartida de las tragedias en este corpus; estas suelen ser conocidos como comedias cómicas (o *urbanas*, o *palatinas*)[10]. De esta manera, el subcorpus anotado contiene en total treinta obras manualmente clasificadas, con ambos géneros representados por la misma cantidad de obras. Las otras ochenta y dos comedias calderonianas están disponibles como textos digitales completos en español modernizado y normalizado[11]. Los parlamentos de los personajes dramáticos se extrajeron de las 112 obras y se recopilaron para su análisis. En los archivos para el análisis no se incluyeron las instrucciones escénicas ni paratextos adicionales. Las quince tragedias se marcaron con una T y un número consecutivo, las comedias con una C y su numeración, y las ochenta y dos obras restantes se marcaron como Test y también se numeraron[12].

2.3. Objetivo de la investigación

La clasificación de género mediante *word* o *document embeddings* es todavía relativamente nueva[13]. Por ello, el objetivo de nuestro estudio es explorar varios métodos y sus combinaciones, y comparar sus resultados. Aplicamos cuatro enfoques que siguen todos los mismos esquemas generales y que serán explicados en detalle en la siguiente sección:

1. Prefiltrado del vocabulario.
2. Cálculo de *document embeddings* y, en su caso, reducción de dimensionalidad.
3. Agrupamiento de *embeddings*.
4. Visualización y evaluación.

Nuestro corpus nos proporciona una base excelente, ya que el género literario se conoce en aproximadamente una cuarta parte de las obras, pero no en el resto de los dramas. De este modo, podemos evaluar simultáneamente la calidad del proceso mediante los textos anotados (con base en las categorías conocidas) y obtener conclusiones sobre los dramas aún no clasificados. Este tipo de comparación metódica nos parece importante, porque se sabe que los resultados de los métodos distributivos no supervisados dependen en gran medida de la parametrización del proceso (Turney & Pantel, 2010; Bullinaria & Levy, 2007).

2.4. Aplicación práctica

Todos los análisis se llevaron a cabo con paquetes del lenguaje de programación R. El preprocesamiento de los textos se realizó en su mayor parte con el paquete R *quanteda*. Este paquete

[10] Véase el resumen más reciente de esta clasificación en Kroll (2022, pp. 63-65). Ver también Calderón de la Barca (1951); Escudero Baztán (2021); Ehrlicher (2012); Maestro (2003); Parker (1988); Peña-Pimentel (2011); Tobar (2000); Valbuena Prat (1950).

[11] En su mayoría, estos dramas están disponibles en el portal Cervantes Virtual http://www.cervantesvirtual.com/ y la Asociación para el Teatro Clásico Hispano http://www.comedias.org/. Un resumen actual de todas las fuentes se puede encontrar en: Estilometría aplicada al Teatro del Siglo de Oro http://etso.es/. Dado que los símbolos diacríticos utilizados en el español moderno pueden emplearse según el contexto, la ortografía de ciertos términos puede variar (p. ej., solos / solós).

[12] Ver el dataset en Zenodo (Lehman, 2022) en el que se ha eliminado esta abreviatura y se presentan los resultados de los métodos aplicados.

[13] Una excepción la constituye el estudio de Willand & Reiter (2017, pp. 190-194).

permite tanto excluir las *stopwords*[14] en español, los signos de puntuación y los números, así como convertir el corpus de textos para que pueda ser procesado por otros paquetes de R. La exploración de los datos muestra que solo un pequeño número de *stopwords* españolas (a saber, 308) fueron eliminadas por el paquete quanteda. Un análisis comparativo mostró que la exclusión de estas *stopwords* de las matrices no conducía a resultados significativamente diferentes, por lo que la lista de *stopwords* se amplió considerablemente de forma manual[15]. Además, el análisis de la estadística numérica tf-idf mostró que los resultados de la agrupación se veían afectados de forma bastante negativa por los nombres de personajes, lugares y países dentro del texto, así como de su forma adjetivada. Esto se debe a que estos elementos del discurso tienden a reflejar la idiosincrasia de cada una de las piezas más que las características del género. Estos nombres propios fueron igualmente identificados y eliminados de los textos, principalmente a través de la lista del *dramatis personae*. Así, a las 308 *stopwords* contenidas en el paquete *quanteda* se le añadieron 800 más.

El siguiente paso consistió en el cálculo de la frecuencia relativa de las palabras en cada drama (normalización de las frecuencias por documento). Posteriormente, se generaron las matrices de distancia y similitud para la agrupación. Al calcular la similitud entre los documentos mediante la similitud del coseno, este paso pudo omitirse, ya que permanecen constantes en relación con las longitudes de los vectores. A lo largo de todos los análisis se trabajó con las formas originales de las palabras (flexionadas o conjugadas), es decir, no se llevó a cabo lematización o stemming.

3. RESULTADOS

3.1. Experimento 0

En una primera exploración, aplicamos un método bien establecido, *skipgram* (Mikolov et al., 2013), al cuerpo del texto para evaluar si las *word embeddings* podían decirnos algo interesante sobre este y qué pares de palabras dentro de todo el corpus de 112 dramas exhibían la mayor cantidad de similitudes. Para ello redujimos la matriz a los 1.000 términos con las probabilidades logarítmicas (*log-likelihood*) más altas y calculamos la similitud del coseno entre todos los pares de vectores. La similitud del coseno o, más exactamente, el coseno del ángulo entre dos vectores es una medida de similitud ampliamente utilizada que determina hasta qué punto dos vectores *apuntan* en la misma dirección en el espacio de alta dimensión. Sus posibles valores están entre 0 y 1 y un coseno alto indica que dos términos se encuentran en contextos similares.

Los pares de palabras con un valor de similitud del coseno muy alto, superior a 0,75, son, por ejemplo, *cielo* y *muerte*, *esperanza* y *desdichas*, *poder* y *temor*, *poder* y *gusto*, *honor* y *alma* o *alma* y *muerte*. Uno de los valores más altos de similitud del coseno, de 0,96, mostró que el par de palabras *honor* y *muerte* puede determinarse como un tema principal en todo el conjunto de la obra. De hecho, estos primeros resultados son sorprendentemente claros, ya que el algoritmo de

[14] Es decir, palabras muy frecuentes de categorías gramaticales como preposiciones, artículos o conjunciones que son borrados para muchos procesos computacionales.
[15] Estas listas de palabras están documentadas en el código R, que fue publicado junto con el cuerpo de dramas en Zenodo en https://doi.org/10.5281/zenodo.6669603. Véase Lehmann, 2022.

skipgram consigue identificar los temas centrales de las comedias calderonianas, incluso cuando tratan de la intersección de las convenciones sociales (honor) y la individualidad (gusto, alma, miedo, muerte social o real).

Por el contrario, emparejamientos de palabras como *honor* y *poder* (0,58), *amores* y *agravios* (0,69) *gracia* y *corte* (0,63) o *gracia* y *culpa* (0,60) mostraron valores de similitud del coseno menores. Los valores de similitud del coseno por debajo de 0,5 muestran aspectos solo poco desarrollados en los textos; esto se pudo observar para los pares de palabras amar y honra, muere y sepulcro, muerte y engaño, mueren y suerte, amores y honra y también mentira y gracia. En primer lugar, se observa que los temas centrales de las obras de Calderón, *amor*, *honor* y *poder* (Escudero Baztán, 2021) no tienen por qué estar interconectados entre sí. Esto debe atribuirse al hecho de que las comedias y las tragedias pueden distinguirse entre sí mediante diferentes combinaciones de estos términos. Es de esperar que la combinación *honor* y *poder* sea más característica de las tragedias, y que la combinación *amar* y *honra* sea más característica de las comedias, pero no del conjunto de las obras. Volveremos sobre este punto más adelante.

3.2. Experimento 1

Con el primer experimento, nuestro objetivo era poder explorar la validez de las *word embeddings*; lo hicimos agrupando los textos para observar si los clústeres muestran los dos géneros de los dramas. Utilizamos las tragedias y comedias anotadas para evaluar nuestros agrupamientos de documentos siguiendo el análisis de la pureza de los clústeres (Manning et al., 2008): asignamos cada clúster a la clase a la que pertenece la mayoría de los documentos con afiliación conocida. A continuación, consideramos las otras clases conocidas de documentos en este clúster, y calculamos la *pureza*, es decir, el grado de acuerdo entre estas clases y la clase mayoritaria, como medida del éxito de nuestra agrupación. Nuestra configuración tiene el aspecto adicional de que nuestro conjunto de datos incluye documentos cuya clase *verdadera* es desconocida. Dado que la pureza solo tiene en cuenta los documentos con clases conocidas, esto hace que la medida sea difícil de interpretar para los clústeres formados predominantemente o en su totalidad por dichos documentos. En el caso de estos conglomerados, que denominamos infradeterminados, nos abstenemos de analizar la pureza en detalle. Tras realizar los pasos de preprocesamiento descritos anteriormente, exploramos los siguientes cuatro métodos:

1. Reducción de la matriz mediante la eliminación de palabras según su frecuencia y aparición en los textos; cálculo de la matriz de distancia según las frecuencias relativas, agrupación con el algoritmo Ward.D2 (Ward, 1963) basado en la distancia euclidiana.
2. Reducción de la matriz mediante la supresión de los *términos dispersos* (es decir, aquellos que solo aparecen en unos pocos documentos), cálculo de la matriz de distancia según las frecuencias relativas, agrupación basada en la distancia euclidiana con el algoritmo de distancia Ward.D2.
3. Anotación automática de las categorías gramaticales (en inglés, *part-of-speech tagging*) en cada uno de los dramas, extracción de verbos, sustantivos y adjetivos, cálculo de los valores de similitud del coseno entre los documentos, cálculo de la matriz de distancia, agrupación con el algoritmo de distancia Ward.D2.

4. Cálculo del algoritmo tf-idf, cálculo de los valores de similitud del coseno entre los documentos, cálculo de la matriz de distancia y agrupamiento con el algoritmo de distancia Ward.D2.

A continuación, mostramos y discutimos los resultados de cada método. El primero representa en los pasos que se realizaron un enfoque conservador: solo se incluyeron las 1.094 palabras con una frecuencia mayor que 120 y que aparecían en al menos la mitad de los documentos. La matriz de palabras del documento se rellenó con frecuencias totales; no se llevó a cabo ninguna reducción de dimensión. El agrupamiento se realizó mediante el algoritmo de distancia Ward.D2. La Figura 1 muestra el dendrograma resultante. Recordemos que entre los documentos que forman los nodos de las hojas del dendrograma, algunos habían sido anotados como comedias (CXX), otros como tragedias (TXX), pero el género de la mayoría es desconocido (Test).

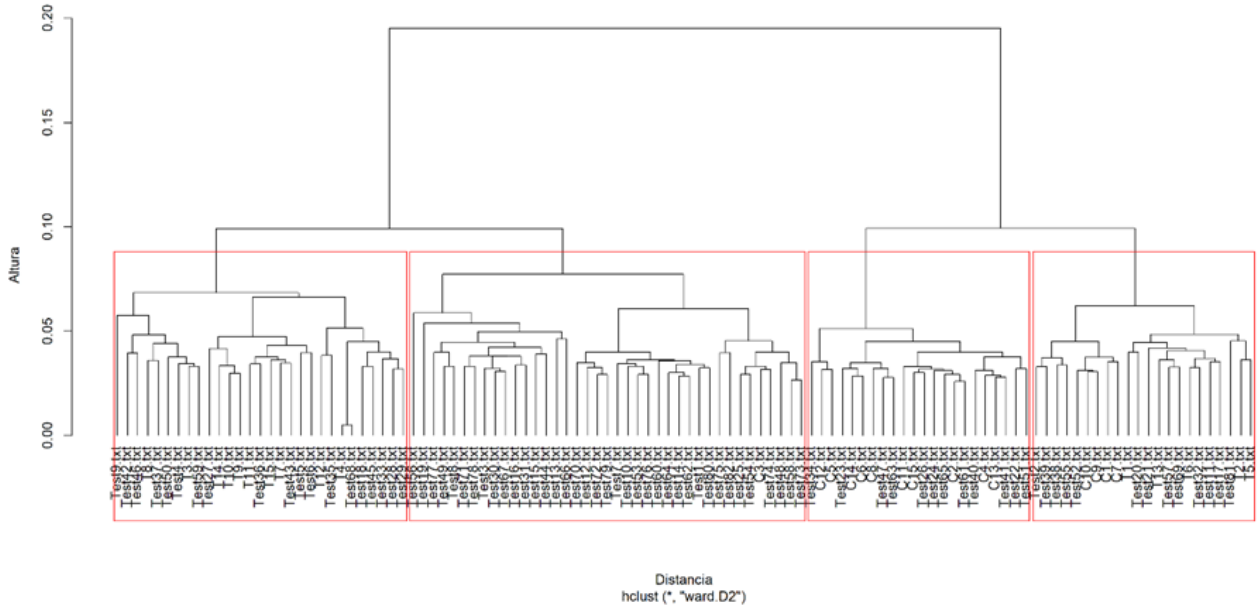

Figura 1. Agrupamiento usando Ward.D2 de 112 comedias calderonianas. Distancia euclidiana basada en la frecuencia normalizada de los *tokens*. Fuente: elaboración propia.

Leído de izquierda a derecha, el primer clúster (resaltado mediante líneas rojas añadidas para mayor claridad) representa un grupo de tragedia pura que incluye 29 dramas; diez de ellos ya habían sido caracterizados como tragedias. El tercer clúster por la izquierda representa un grupo de comedia pura; aquí se incluyen 22 dramas, de los cuales diez ya habían sido clasificados como tragedias. Los otros dos clústeres (segundo y cuarto) los consideramos clústeres mal definidos o mixtos, ya que o bien contienen una sola comedia (segundo clúster por la izquierda, que incluye 39 dramas), o bien contienen cuatro comedias y cinco tragedias (el clúster de la derecha, que incluye 22 dramas). En conjunto, estos dos clústeres contienen más de la mitad de las obras, 61 en concreto. Llegamos a la conclusión de que este enfoque, con respecto a nuestra principal pregunta de investigación, no parece organizar las obras según su género de manera especialmente eficaz, ya que solo 20 de los 30 dramas anotadas previamente (el 67 %) se agruparon de forma clara, mientras que las diez comedias y tragedias restantes aparecieron en los clústeres mixtos. Sin embargo, la dimensionalidad aún relativamente alta de los *word embeddings* dificulta su análisis.

El objetivo del segundo proceso es crear una representación de baja dimensión que sea más fácil de interpretar, para obtener más información sobre la distribución de los dos géneros. En

primer lugar, solo se conservan los términos que aparecen en más del 80 % de los documentos (es decir, los términos que aparecen en por lo menos 90 obras). De esta manera, la dispersión se limita al 20 %. Esto reduce el número de términos a un total más reducido de 496 palabras. De nuevo, se establece una matriz palabra-documento basada en la frecuencia y se normaliza su frecuencia, mediante la división de cada uno de los términos por la suma de las frecuencias de todas las palabras. Por último, se establece una matriz de distancia, basada en la distancia euclidiana y, de nuevo, se realiza la agrupación mediante el algoritmo de distancia Ward.D2.

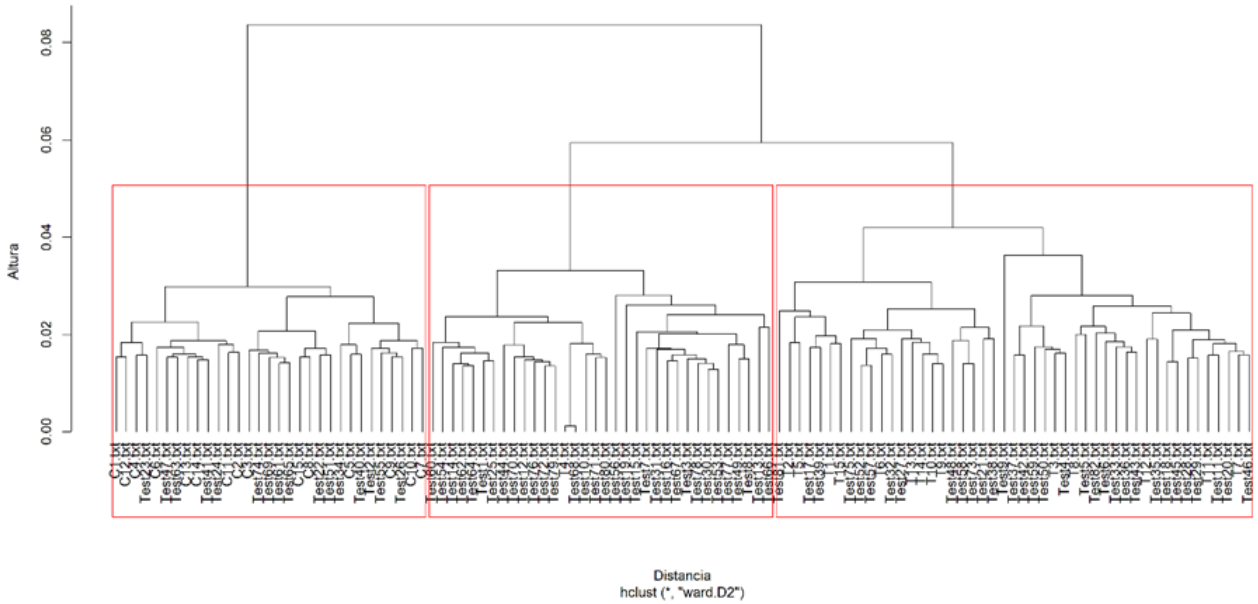

Figura 2. Agrupamiento usando Ward.D2 de 112 comedias calderonianas. Distancia euclidiana según una dispersión (sparsity) del 20 %. Fuente: elaboración propia.

En este caso, el dendrograma muestra tres clústeres principales (de nuevo marcados en rojo por nosotros): en el primero por la izquierda aparecen las 15 comedias y otros 16 dramas sin anotación de género. El clúster de la derecha contiene 14 tragedias y 33 dramas de clasificación desconocida. El clúster del medio resulta mixto, ya que contiene una tragedia (T4: *El mayor monstruo del mundo*) y 33 dramas adicionales de clasificación desconocida. A través de este proceso, que solo considera 496 palabras, se asignaron correctamente 29 de los 30 dramas clasificados, es decir, el 97 %[16].

Los dos métodos de agrupamiento utilizados hasta ahora, en los que las matrices originales se reducen a partir de las frecuencias de las palabras, establecen un clúster de transición entre la tragedia y la comedia. Esta observación nos plantea la cuestión de si no sería más adecuado, a la luz de la semántica distribucional, considerar como *tragedia* y *comedia* como clases, entre los que aparecen diferencias graduales, mostrando el solapamiento resultante respecto a la selección de palabras aplicada. En el caso de los dramas calderonianos, esto parece bastante sensato, ya que temas como el honor y el poder pueden estar incluidos tanto en las tramas cómicas como en las de

[16] Básicamente, intentamos alterar solo un parámetro entre cada uno de los cuatro análisis presentados en esta sección. Por esta razón, en este segundo experimento mantenemos la distancia euclidiana. Sin embargo, también utilizamos la distancia Manhattan durante el segundo experimento, que se define como la distancia por la suma de valores absolutos. Los resultados fueron claramente menos satisfactorios que las representaciones anteriores resultantes del uso de la distancia euclidiana: solo dos tercios (67 %) de todas las tragedias y comedias previamente identificadas fueron clasificadas correctamente.

las famosas tragedias de honor.

Las comedias también pueden presentar temas serios de forma desenfadada y divertida. Por ejemplo, se puede aludir indirectamente a las luchas de poder entre las familias reales en el marco de una obra mitológica; la alegoría habría sido bastante comprensible para el público de la corte de la época[17].

Una posible crítica fundamental a los métodos simples de *word embeddings* como los que hemos visto hasta ahora es la ausencia total de estructura lingüística. Por esta razón, tomamos la decisión de anotar en todos los dramas las categorías gramaticales de cada palabra (*part of speech*). Posteriormente se eliminaron las palabras marcadas con categorías diferentes a verbos, sustantivos y adjetivos de cada obra. La matriz usada para el agrupamiento se creó así con palabras pertenecientes a estas categorías (Willand & Reiter, 2017, pp. 191 y ss.). Así, para probar el tercer procedimiento, se establece una segunda versión del corpus en el que cada uno de los textos dramáticos incluye solo verbos, sustantivos y adjetivos en sus formas básicas lematizadas. Todos los nombres propios se eliminaron de la matriz creada a tal efecto, ya que se habían reconocido erróneamente como adjetivos. Posteriormente, se realiza el cálculo de las frecuencias no normalizadas mediante la similitud del coseno. Esta matriz de similitudes se convierte en una matriz de distancias y, de nuevo, se clasifica con el algoritmo Ward.D2. Los resultados se representan en un dendrograma de la Figura 3.

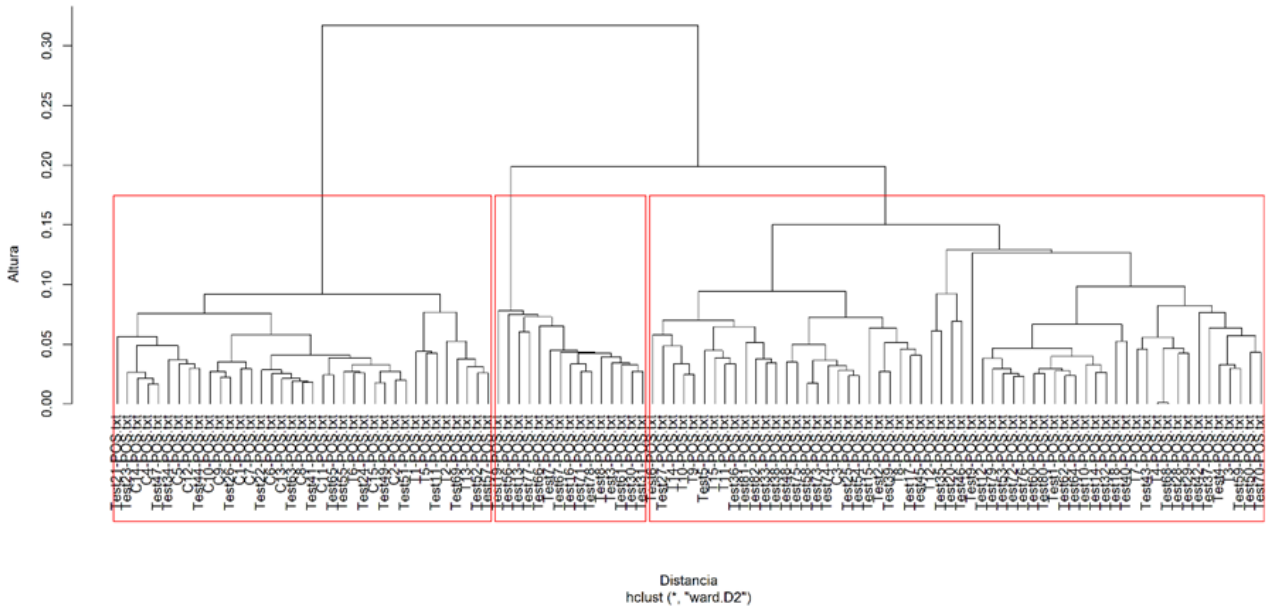

Figura 3. Agrupamiento usando Ward.D2 de 112 comedias calderonianas. Similitud de coseno basada en verbos, sustantivos y adjetivos. Fuente: elaboración propia.

El primer clúster por la izquierda, que podría identificarse como un grupo de comedia, contiene catorce comedias, cinco tragedias (T1: *A secreto agravio, secreta venganza*; T2: *El alcalde de Zalamea*; T5: *El médico de su honra*[18]: T6: *El pintor de su deshonra*; T13: *Las tres justicias en una*) y 18 obras adicionales cuyo género no ha sido todavía anotado. El grupo de la derecha es principal-

[17] Esta posibilidad ya fue mencionada por Greer (1988) en un ejemplo de *Fieras afemina Amor*.

[18] Este resultado es especialmente interesante porque, según Couderc (2012, p. 104), *A secreto agravio, secreta venganza* y *El médico de su honra* pueden calificarse de tragicomedias y *A secreto agravio, secreta venganza* es la única obra de Calderón que utiliza el término "tragicomedia" en el texto hablado.

mente un clúster de tragedias que contiene diez obras de este género y 49 obras no anotadas, así como una comedia (C3: El encanto sin encanto). En medio de estas dos categorías se encuentra un clúster con 15 obras sin anotación alguna de género (*Test*). Con respecto a las obras anotadas hasta ahora como tragedias o comedias, el 80 % de los dramas fueron clasificados correctamente; sin embargo, este resultado solo se aplica si los clústeres son identificados por la mayoría de los dramas previamente identificados[19]. Teniendo en cuenta los métodos probados anteriormente, resulta obvio que las matrices utilizadas en los dos primeros procedimientos contenían más información lingüística porque estaban compuestas por preposiciones, determinantes y conjunciones; los clústeres resultantes eran a menudo puros. Por lo tanto, parece aconsejable centrarse en todos los términos anotados como verbo, sustantivo o adjetivo y que, por lo tanto, conducen a una diferenciación entre las categorías.

El cuarto método que hemos probado se basa en la estadística numérica tf-idf, una medida de asociación comúnmente utilizada en *minería de textos*, mediante la cual se puede evaluar la importancia de los términos dentro del documento o del corpus. Con tf-idf se calcula el peso de cada término por documento; la *frecuencia del término* (*tf*) se multiplica por la *frecuencia inversa del documento* (*idf*). Esta última no depende de los documentos individuales, sino de la frecuencia total en todos los documentos del corpus. De este modo, la estadística numérica tf-idf tiene en cuenta la importancia relativa de las palabras que aparecen con frecuencia en el corpus para determinar la relevancia del término en un documento dentro del corpus analizado. Una vez más, se eliminan los nombres propios (véase explicación en el anterior método), se calcula la similitud del coseno para los vectores, se convierte la matriz de similitud en una matriz de distancia y se realiza la agrupación con el algoritmo Ward.D2. Los resultados se representan en un dendrograma.

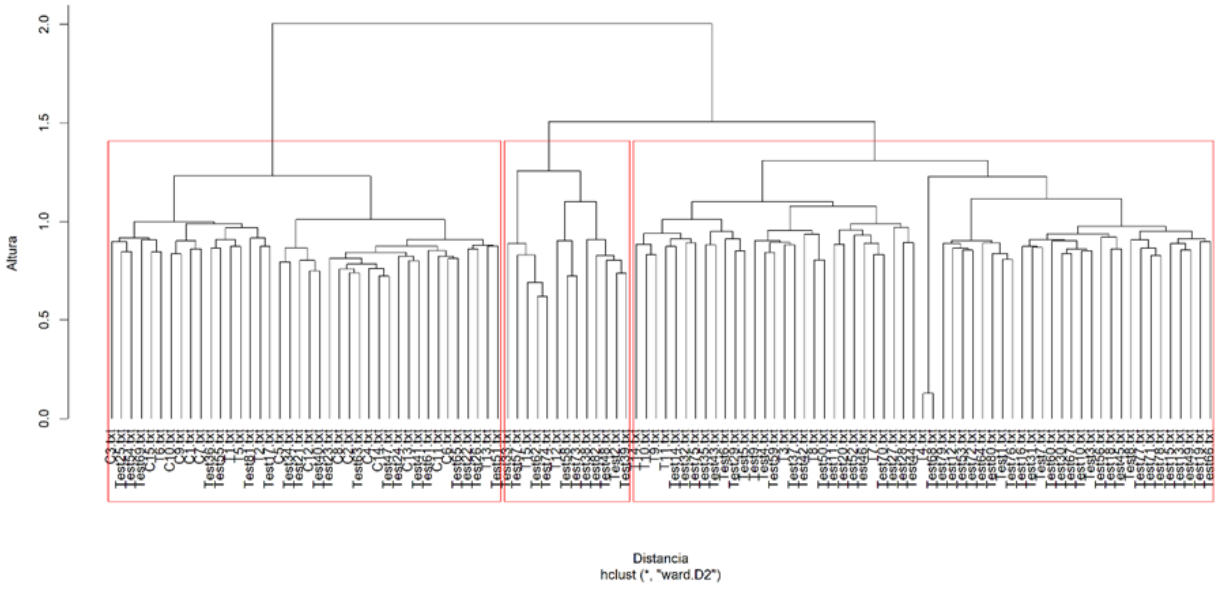

Figura 4. Agrupamiento Ward.D2 de 112 comedias calderonianas. Similitud del coseno sobre la base de los valores tf-idf. Fuente: elaboración propia.

[19] Como alternativa, se estableció una matriz normalizada y se realizó un agrupamiento Ward.D2 basado en la distancia euclidiana. Los resultados son más claros, ya que cuatro tragedias y catorce comedias fueron asignadas a un clúster no mixto. Sin embargo, las once tragedias restantes y una comedia formaron un clúster mixto, por lo que en total solo se alcanzó una pureza del 60 % en la clasificación. Compárese con el script R de la publicación de datos y código.

La Figura 4 muestra tres clústeres: el primero por la izquierda podría describirse como un clúster de comedias. Sin embargo, además de quince comedias, contiene también cinco tragedias, exactamente las mismas cinco que en el análisis basado en partes del discurso realizado anteriormente (T1, T2, T5, T6, T13), así como otros veinte dramas sin anotación del género. El clúster de la derecha, con ocho tragedias y cincuenta y una obras más, puede considerarse un clúster de tragedias. El más pequeño, situado en el centro, resulta difícil de definir ya que es infradeterminado y solo contiene dos dramas claramente identificados como tragedias y otros once que permanecen sin clasificar. En comparación con los dramas ya identificados como tragedias o comedias, este resultado muestra que ocho de quince tragedias y todas las comedias han sido clasificadas correctamente. Esto representa una tasa de reconocimiento del 76 %[20]. Si se compara con los modelos anteriores, esta tasa parece satisfactoria.

Los cuatro métodos aplicados aquí difieren en la elección de los datos, así como en su transformación y la elección de las métricas de distancia o similitud. Tres de ellos generaron resultados que podrían considerarse como entre aceptables y muy buenos. El proceso de emplear una reducción de matriz más compleja produjo los mejores resultados. Sin embargo, solo uno de los enfoques produjo un clúster que podría aproximarse a la clasificación producida por los investigadores de manera cualitativa.

3.3. Experimento 2

En este segundo experimento, evaluamos hasta qué punto los clústeres de documentos que encontramos en el primer experimento se basan en una frecuencia de uso de las palabras consistente dentro de cada uno de los géneros. Para ello, analizamos las listas de palabras en las que se basaron las agrupaciones encontradas por los cuatro métodos. Además, calculamos la distribución de *log-likelihood* del vocabulario para los conjuntos de comedias y tragedias (clasificadas) de cada método. De esta manera, determinamos las 200 palabras con los valores más altos de *log-likelihood* para cada género, y estas listas pueden compararse entre métodos (análisis de vocabulario contrastado con *word embeddings*).

Recordemos que el primer procedimiento del primer experimento –agrupamiento de Ward.D2 basado en la distancia euclidiana entre las frecuencias normalizadas de las palabras)– agrupó las obras de forma que solo el primer y el cuarto clúster podían evaluarse claramente como de comedia o de tragedia respectivamente. Para estos dos clústeres, se evalúa el margen de probabilidad de cada palabra a partir de la matriz previamente establecida y se investigan los quince términos con mayor margen de probabilidad para cada uno. Estos quince términos seleccionados para los clústeres de comedia y tragedia con los márgenes de probabilidad más altos dan una idea de la formación de los clústeres. Para el clúster de comedia, aparecieron los términos

[20] Aquí, alternativamente, también se llevó a cabo una agrupación Ward.D2 basada en la distancia euclidiana. El resultado muestra cinco conglomerados, tres de los cuales contienen cuatro dramas etiquetados como Test. Los dos clústeres restantes consisten en un clúster mixto que contiene quince comedias, doce tragedias y cuarenta obras más; y otro clúster que contiene tres tragedias, así como treinta y ocho obras más. Estos resultados confirman la poca fiabilidad de este enfoque en lo que respecta a la clasificación.

don, casa, calle, papel, caballero, puerta, dama, padre, hermano, saber, cuarto, amigo, hombre, sé y *señora*. Muy interesante es la palabra *papel*, ya que señala el papel o la tarjeta que aviva la intriga; sin embargo, más allá de este término, la lista de palabras no parece ser significativamente distintiva de las comedias. En cambio, para el grupo de las tragedias, son especialmente frecuentes las palabras *rey, muerte, dios, cielo, hoy, vida, sol, valor, mar, tierra, gran, rigor, mundo, quiero* y *poder*. En cualquier caso, las personas de alto nivel social, la muerte, Dios, el valor y el poder destacan como términos característicos relacionados con estos argumentos.

Las 496 palabras seleccionadas por su ausencia en el 20 % del corpus permiten una exploración de términos que conllevan una fuerte distinción entre comedias y tragedias. Para el grupo de comedias, están presentes términos significativos como *don, casa, dama, calle, puerta, sé, señor, caballero, bien, cuarto, papel, señora, saber, amigo* y *celos*. En el grupo de la tragedia aparecen palabras como *rey, señor, dios, hoy, muerte, cielo, sol, quiero, rigor, mundo, gran, valor, alma, viento* y *sangre*. Principalmente, puede sorprender el alto grado de consistencia de las listas de palabras del primer y segundo procedimiento. Por otra parte, parece que el alto grado de pureza del agrupamiento en el segundo procedimiento depende obviamente de la selección condensada y precisa de los términos distintos.

En cuanto al tercer procedimiento, basado en un corpus etiquetado con información sobre las categorías gramaticales, las palabras más frecuentes encontradas en los grupos de la matriz subyacente ilustran por qué no conduce a resultados convincentes: no es de extrañar que las palabras más frecuentes sean los verbos *ser* y *haber*, seguidos a una distancia notable por una lista de verbos adicionales, como *ver, decir, estar, dar, poder, saber, hacer, tener, ir, querer, venir*. A continuación, se encuentra una lista de sustantivos, como *señor, vida, cielo* o *don*. Teniendo en cuenta que estas palabras de uso frecuente parecen tener poca capacidad para distinguir entre comedias y tragedias, es de esperar que los resultados de la agrupación pueden calificarse de bastante pobres.

En el cuarto procedimiento (basado en la matriz tf-idf) se aplica un enfoque análogo al primer y segundo método. Los quince términos que muestran el mayor margen de probabilidad dentro del clúster de comedia son: *don, doña, tapada, hermana, calle, hermano, coche, amiga, anoche, papel, cuarto, aposento, reja* y *casa*. En el grupo de la tragedia son característicos términos como *arma, dioses, cristianos, templo, montes, cueva, ciencias, cruz, muro, reino, pastor, rey, cristiano, cajas* y *guerra*. Mientras que los términos frecuentes seleccionados para el grupo de la comedia parecen, en su mayor parte, menos específicos del género –salvo las típicas alusiones al velo y al enmascaramiento o a la intriga por medio de la falsificación–, los términos relativos a la tragedia reflejan al menos temas militares y cristianos, así como la ascendencia aristocrática de los protagonistas.

Una cuestión abierta en este punto es la solidez de estos métodos. Por este motivo, en el siguiente paso ponemos a prueba las listas de palabras creadas en los pasos anteriores basándonos en un corpus de obras más amplio: ampliamos nuestra base de datos a grupos y ampliamos el conjunto de obras identificadas como comedias o tragedias y creamos dos subgrupos algo mayores. De los ochenta y dos dramas sin anotación alguna de género y marcados hasta ahora como

Test, elegimos dieciséis que fueron identificados unánimemente como tragedia por los cuatro procedimientos, así como diez clasificados unánimemente junto a comedias. Para estas, corroboramos esta clasificación a partir de la literatura secundaria[21]; además, todos estos dramas habían sido ya incluidos en la colección de comedias por los editores de la edición de Aguilar. De este modo, generamos dos nuevos subcorpus: uno con treinta y una tragedias, y otro con veinticinco comedias (ver para un método comparativo Peirsman et al., 2010). Ambos subgrupos se convierten en matrices utilizando las técnicas de preprocesamiento habituales, por las que se filtran todos los términos que se encuentran en menos de cuatro de las obras. Para las palabras restantes, se identifican las 200 más informativas de cada subgrupo para su inclusión utilizando la función de *log-likelihood*, con la que se pueden encontrar términos discriminadores. La comparación de los resultados de cada subgrupo muestra que solo setenta términos aparecen en ambas listas, mientras que 130 términos de cada una (casi dos tercios) son discriminadores, ya sea para el subgrupo de tragedia o para el de comedia.

El análisis de estos 130 términos discriminadores para cada subgrupo resulta muy revelador. En el caso de las comedias, descubrimos referencias a determinados temas (*ama, amiga, carta, celoso, desdichas, desengaño, escondido, favor, joyas, juego, máscara, papeles, secreto, tapada, vestido*), indicaciones típicas relacionadas con el trasfondo mitológico de las comedias (*astrólogo, duende, forastero, jardines, ninfas*) y también la aparición de algunos términos bastante sorprendentes, (como *enemigo, pendencia, razón,* o *saber*).

En cambio, en la lista de términos de las tragedias encontramos referencias a la posición (en su mayoría alta) de los personajes (*convento, corona, emperador, esclavo, infanta, infante, majestad, reina, reinar, reino, rey, tirano, villano*), el contenido de la trama (*cristo, cruz, desdichado, divina, esperanza, gloria, laurel, lealtad, libertad, morir, poder, salud, sangre, traición, triste, triunfo, venganza, victoria*) y también algunos términos sorprendentes (*ciencias, enamorado, sueño*). En conjunto, las listas de palabras que determinan la log-likelihood en los dos géneros esbozan el contenido de las comedias y las tragedias con mucha más precisión que las listas de palabras basadas en cada clúster.

3.4. Experimento 3

En nuestro último experimento, pasamos del análisis de documentos en términos de palabras, como en el Experimento 2, a un análisis del uso de palabras individuales en los dos géneros. Para ello, utilizamos el método de *embeddings fastText* (Bojanowski et al., 2017) y el paquete R del mismo nombre. A diferencia de *skipgram*, *fastText* es más apropiado para cuerpos de texto más pequeños, ya que no calcula un *embedding* para cada palabra. En su lugar, se calculan los

[21] Casi todos estos dramas entran en la categoría de *comedias cómicas* descrita por Kroll (2022, pp. 64-65). Sin embargo, hay dos excepciones: en contraste con la estimación de Kroll, que sitúa *No hay cosa como callar* en la categoría *tragedias y dramas de honor*, nosotros clasificamos este drama como *comedia*, ya que los cuatro métodos empleados estaban de acuerdo. En comparación, descartamos *Las manos blancas no ofenden* de la lista de comedias, ya que la estimación de Valbuena Prat (1950, p. 451), que cuenta esta obra entre las "obras exclusivamente cómicas", no fue corroborada por los procedimientos que aplicamos.

embeddings de partes de palabras (por ejemplo, para honor: *hon, ono, nor*, etc.) y se acumulan para crear un *embedding* de la palabra completa. De este modo, surgen representaciones más sólidas para las palabras poco utilizadas o desconocidas (Papay et al., 2018). En cada subgrupo, se establecen los 10 términos vecinos de interés más cercanos dentro del *word embeddings*, de modo que cada palabra que se identificó como distintiva de ambos géneros es visible, junto con los términos que suelen coocurrir más frecuentemente en los textos.

Para contrastar los términos de cada subcorpus, a continuación, ilustraremos los diez términos más próximos por subcorpus junto con las similitudes de cada uno, donde la máxima similitud posible tiene valor 1.

La palabra clave *honor*, que se encuentra no solo en las comedias, sino también en las tragedias, no muestra términos vecinos comunes en el subgrupo de las tragedias cuando se evalúa dentro del subgrupo de las comedias, ni tampoco se encontraron estos para la palabra *hado*. En otros términos, ambos términos se utilizan en comedias y tragedias, pero dentro de contextos completamente diferentes según cada una. Se pone de manifiesto que los términos *honor* y *hado* que aparecen en las tragedias están más claramente perfilados dentro del contexto y, teniendo este en cuenta, el significado de los términos puede entenderse con mayor precisión. Por ejemplo, el *honor*, dentro del contexto de la tragedia, se refiere a la pérdida del mismo, o a la difamación, cuyo re-

Comedia	Tragedia
honor	
pundonor 0,81	satisfación 0,81
ofrecer 0,80	sujeción 0,78
lograr 0,79	oración 0,77
honrar 0,79	rigor 0,76
obedecer 0,78	maldición 0,76
menor 0,78	opinión 0,75
reconocer 0,78	satisfecha 0,75
rencor 0,77	satisfacción 0,75
confesar 0,77	honra 0,75
ofender 0,77	acción 0,75
hado	
hallado 0,92	estimado 0,92
amado 0,91	librado 0,91
hablado 0,91	engañado 0,90
madrugado 0,90	sobrado 0,88
echado 0,90	nombrado 0,88
mirado 0,89	tratado 0,88
negado 0,89	rendido 0,87
pecado 0,89	desengañado 0,87
tocado 0,87	mostrado 0,87
enfadado 0,87	estrado 0,87

Tabla 1. 10 términos vecinos más cercanos para *honor* y *hado*. Fuente: elaboración propia.

medio está obviamente asociado a la posible muerte.

Que muchas palabras terminen de manera similar en esta tabla puede resultar desconcertante a primera vista, pero no debería ser sorprendente: todas las obras de Calderón están escritas en verso. Solo por la métrica, la selección de posibles palabras vecinas está drásticamente limitada. Para limitarlo aún más, las inflexiones y conjugaciones similares de la lengua española dejaron a Calderón con una selección reducida de posibles palabras al componer las oraciones de sus obras dramáticas.

Otros términos que se utilizaron en ambos subgrupos también producen resultados similares a *honor* o *hado*. Las palabras *fineza*, *justicia* y *amistad* solo arrojaron una o dos palabras vecinas comunes en ambos subgrupos (representadas en negrita en la siguiente tabla); estos términos se encuentran tanto en las comedias como en las tragedias, pero en contextos muy diferentes. Mientras que estos tres términos en el contexto cómico tienden a reflejar lo profano, su aparición en el contexto trágico refleja la autoridad formal del tribunal y su jurisdicción, así como la seriedad y el ámbito de la providencia y la justicia divinas.

Comedia	Tragedia
fineza	
firmeza 0,84	fiereza 0,84
fianza 0,81	gloria 0,78
importuna 0,81	peregrina 0,77
fina 0,80	indignación 0,77
impida 0,80	insignia 0,77
implica 0,79	ofrecí 0,76
naturaleza 0,79	grandeza 0,76
nobleza 0,78	**firmeza** 0,75
templanza 0,78	imperial 0,75
belleza 0,77	ignorancia 0,75
justicia	
justa 0,83	**justa** 0,83
hidalga 0,78	justiciero 0,82
acompañada 0,77	licencia 0,80
malicia 0,77	precia 0,79
salida 0,76	milicia 0,79
diligencia 0,75	**malicia** 0,78
hidalguía 0,75	usted 0,77
historia 0,75	gusta 0,77
dispensación 0,75	estudiar 0,77
traición 0,75	condición 0,76

Comedia	Tragedia
amistad	
dad 0,85	acudid 0,82
vanidad 0,83	**calidad** 0,82
mitad 0,83	ofrezca 0,81
debéis 0,83	seguridad 0,81
decid 0,81	fealdad 0,77
calidad 0,81	temeridad 0,77
mirad 0,80	**mitad** 0,77
libertad 0,80	sacad 0,76
perdonad 0,79	firmeza 0,76
podáis 0,79	salid 0,76

Tabla 2. 10 términos vecinos más cercanos para *fineza*, *justicia* y *amistad*. Fuente: elaboración propia.

Sin embargo, otros términos muestran claramente solapamientos con respecto a los términos

vecinos más cercanos; por ejemplo, celos, gusto o muera comparten cada uno tres o cuatro términos vecinos más cercanos dentro de las diez palabras de la selección.

Comedia	Tragedia
celos	
celosos 0,91	consuelos 0,91
recelos 0,90	**recelos** 0,91
duelos 0,89	**celosos** 0,90
cielos 0,85	antojos 0,89
puestos 0,84	pueblos 0,89
palos 0,83	regalos 0,88
dellos 0,83	demos 0,88
desconsuelos 0,82	**cielos** 0,87
opuestos 0,82	caballos 0,87
laberintos 0,82	verlos 0,87
gusto	
admito 0,87	**justo** 0,87
visto 0,86	desprecio 0,85
susto 0,86	precio 0,84
justo 0,84	justiciero 0,84
gasto 0,84	**disgusto** 0,83
disgusto 0,84	precepto 0,82
pedido 0,83	preciso 0,82
considero 0,82	profano 0,82
adentro 0,82	favorecido 0,82
pecado 0,82	convencido 0,82
muera	
muriera 0,89	viviera 0,94
muerta 0,89	**muriera** 0,94
defuera 0,85	**muerta** 0,92
muralla 0,85	**muralla** 0,91
muestra 0,84	diera 0,90
manera 0,83	madera 0,90
mira 0,82	**manera** 0,90
enferma 0,81	viera 0,90
dondequiera 0,81	hermosura 0,89
cólera 0,81	matara 0,89

Tabla 3. 10 términos vecinos más cercanos para *celos*, *gusto* y *muera*. Fuente: elaboración propia.

Este análisis ilustra que las diferencias entre las tragedias y las comedias no consisten simplemente en un vocabulario diferente: incluso el léxico compartido se utiliza sustancialmente de forma distinta. Cuanto más central sea el término para el género, más se distingue el uso entre ambas categorías. Al menos, esta es la tendencia que han mostrado nuestros resultados hasta ahora.

4. DISCUSIÓN DE LOS RESULTADOS, CONCLUSIONES Y SIGUIENTES PASOS

La comparación de los métodos muestra que con dos de ellos (la clasificación de dramas usando las frecuencias de verbos, sustantivos y adjetivos, y usando los valores tf-idf) se pueden alcanzar resultados que se aproximan a la clasificación de los expertos. Ambos métodos se consi-

deran procedimientos estándar en la *minería de textos*. Sin embargo, para que la clasificación alcance una pureza de al menos el 67 %, es necesario un filtrado exhaustivo que vaya más allá de los signos de puntuación y las palabras vacías habituales y que incluya otras palabras funcionales, nombres propios y sus formas adjetivadas. Estas últimas tienen que ser tratadas en parte manualmente para cada corpus estudiado, lo que requiere un esfuerzo considerable. Se puede alcanzar una pureza de clasificación bastante buena con cierta rapidez si se lleva a cabo una reducción masiva de la matriz de rasgos, eliminando los términos con dispersión menor del 20 %, es decir, considerando así solo los términos que aparecen en al menos el 80 % de todos los documentos.

Las observaciones preliminares de este estudio considerando la comparación de los cuatro métodos explorados permiten identificar con una alta probabilidad más dramas de cada categoría, en concreto dieciséis tragedias y diez comedias. Los resultados también señalan que el vocabulario utilizado para caracterizar cada una de ambas categorías comparte términos con el otro, así como ciertos resultados contradictorios. Esto afecta especialmente a los pasajes cómicos de los dramas (incluso cuando aparecen dentro de una tragedia), pero también a los términos que reflejan temas típicos de comedias o tragedias, atributos extraliterarios o características de la trama.

Un ejemplo particular sería la obra *Amor, honor y poder*, un título desconocido para los autores de este estudio antes de comenzar el análisis. Aunque comúnmente se clasifica como comedia por su final feliz, la intriga trata de las relaciones infelices entre dos parejas de personajes y, por tanto, está dominada por una semántica propia de las tragedias. Otra excepción la constituye *No hay cosa como callar*, obra que, de nuevo, los cuatro procedimientos clasifican de forma unánime junto con las comedias, como también lo hace la edición de Aguilar. Los juicios de la investigación cualitativa, sin embargo, están más divididos: mientras que Alexander A. Parker (1962, p. 228) lo clasificó en 1962 como tragedia, posteriormente revisó su juicio y lo calificó de "comedia de intriga" (Parker, 1988, pp. 181-182), y Simon Kroll (2022, p. 63) lo sitúa en el apartado de "tragedias y dramas de honor". Sin duda, el análisis realizado aquí inspirará nuevos debates, ya que tales diferencias en la clasificación de un drama pueden resolverse mediante un examen diferenciado: el vocabulario de *No hay cosa como callar* puede ser el típico de las comedias, pero la trama, así como otros criterios cualitativos, podrían apoyar su clasificación como tragedia.

También es interesante la constatación de que las tragedias calderonianas son mucho más identificables que las comedias, por la forma en que se utilizan las palabras dentro del texto. Esto queda subrayado por la manera en la que los cuatro métodos aplicados agruparon el grupo de las llamadas *comedias religiosas*: *El José de las mujeres, El purgatorio de san Patricio, Judas Macabeo, La cisma de Ingalaterra, La exaltación de la cruz, La sibila del Oriente y gran reina de Sabá, Las cadenas del demonio, Los dos amantes del cielo*, y *Origen, pérdida y restauración de la Virgen del Sagrario*. Todos estos dramas se caracterizan por el uso de un vocabulario trágico. Por otra parte, en lo que respecta a las comedias, es evidente que son mucho más difíciles de definir que las tragedias. Esto es cierto, por ejemplo, con respecto a un grupo de comedias que frecuentemente se consideran comedias mitológicas, como *El castillo de Lindabridis, El mayor encanto, amor, La puente de Mantible* y *Los tres mayores prodigios*, que muestran señales de tragedia muy fuertes en nuestro

análisis, mientras que la mayoría de los otros dramas clasificados como comedias mitológicas muestran señales mixtas[22].

Ciertamente, la separación binaria de *dramas* y *comedias* realizada anteriormente por los editores de la edición de Aguilar debe ser revisada de manera crítica. Esto es especialmente claro en aquellos dramas que hasta ahora han recibido poca atención. Un buen ejemplo de ello lo proporciona *Amar después de la muerte*, que destaca por el uso del vocabulario trágico, tal y como lo identificó el enfoque de clasificación más puro (método 2). Esta clasificación fue verificada por la edición histórico-crítica presentada por Jorge Checa (2010). Dado que este, en el prefacio de su análisis, discute una serie de criterios relativos a la designación de las tragedias según Parker y Sullivan (Checa, 2010, pp. 12-16), nuestros resultados presentan una invitación a los investigadores que trabajan cualitativamente para hacerlo de forma sistemática y aplicar de forma coherente estos criterios establecidos para la clasificación en una serie completa de obras. El estatus del grupo de dramas llamados comedias mitológicas (al igual que el de los reconocidos por Parker y Sullivan como "al borde de la tragedia")[23] debe, por lo tanto, ser discutido de nuevo con respecto a sus categorías designadas y al vocabulario utilizado. Lo mismo ocurre con el grupo de dramas escasamente examinados que pueden clasificarse como tragicomedias. La zona intermedia que se encuentra entre las comedias y las tragedias a lo largo de estos métodos lo señala de forma rotunda. En consonancia con las Humanidades Digitales, esta conclusión representa una invitación a que los investigadores profundicen en los textos de manera cualitativa y que creen listas de palabras características para cada categoría que se quiera distinguir.

Consideramos que el enfoque realizado a través de la semántica distribucional es una de las posibilidades de clasificar las obras, aunque es solo una de ellas. Este análisis resulta particularmente interesante cuando, como es el caso, los análisis léxicos y semánticos van de la mano. Una ventaja importante de los *word embeddings* en comparación con los modelos de bolsas de palabras (*bag-of-word models*) es que los primeros fundamentan las representaciones de las palabras –y, por tanto, también las representaciones de los documentos– no solo en símbolos, sino en contextos, lo que nos permite comprender de manera más matizada en qué se distinguen los documentos. Esto también permite detectar diferencias por debajo del nivel de los símbolos o de su uso, mostrando, por ejemplo, que ciertas palabras (*honor*, *hado*) se utilizan tanto en las tragedias como en las comedias, pero con colocaciones totalmente diferentes y, por tanto, con significados distintos. Esta idea es especialmente relevante si se tiene en cuenta el gran número de obras que aún no han sido investigadas o no en suficiente profundidad. La comparación sistemática de varios métodos, como la que aquí se han empleado, ofrece la oportunidad de evaluar mejor los resultados de corpus heterogéneos (obras de varios dramaturgos o de diferentes siglos). La aplicación de estos procedimientos a, por ejemplo, todos los dramas del Siglo de Oro disponibles proporcionaría una base más amplia sobre el léxico característico de las comedias y las tragedias. Precisamente, el ejemplo

[22] Las valoraciones de estas obras como *comedias mitológicas* de Kroll (2022); Castro de Moux (2001); Greer (1988); Cancelliere (2000), Arellano (2000), Peña-Pimentel (2011).
[23] Ver Parker (1988, pp. 58, 181, 182); Sullivan (2018, pp. 70, 316, 321).

de Calderón con sus 112 comedias nuevas ilustra que los métodos aquí explorados proporcionan a la comunidad investigadora información que puede estimular otros análisis. Potencialmente, iniciativas actuales que preparan ediciones históricos-críticas de todos los dramas calderonianos[24] pueden utilizar los resultados presentados en este estudio.

REFERENCIAS BIBLIOGRÁFICAS

Arellano, I. (2000). El Teatro de Corte y Calderón. En M. L. Tobar (Ed.), *Atti della Tavola Rotonda sulla Singolarità Storica e Estetica di "La púrpura de la rosa" di Calderón de la Barca* (pp. 31-53). Armando Siciliano Editore.

Arellano, I. (2007). *Editar a Calderón. Hacia una edición crítica de las comedias completas*. Universidad de Navarra.

Arellano, I. (2018). Calderón y los géneros dramáticos, con otras cuestiones anejas. Honor, amor, legitimación política y autoridad de las taxonomías. *Rilce. Revista de Filología Hispánica*, *34* (1), 100-126. https://doi.org/10.15581/008.34.1.100-126

Benjamin, B. (1978). *Ursprung des deutschen Trauerspiels*. Suhrkamp.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135-146. https://aclanthology.org/Q17-1010.pdf

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics. A computational study. *Behavior Research Methods*, *39*, 510-526. https://doi.org/10.3758/BF03193020

Calderón de la Barca, P. (1951-1956). *Obras completas. Textos íntegros según las primeras ediciones y los manuscritos autógrafos*. Ed. Á. Valbuena Briones, & L. Astrana Marín. 3 vols. Aguilar.

Calderón de la Barca, P. (2006-2010). *Comedias y otras obras*. Ed. L. Iglesias Feijoo. 6 vols. Fundación José Antonio de Castro.

Campión Larumbe, M., & Cuéllar, Á (2021). Discernir entre original y refundición en el teatro del Siglo de Oro a través de la estilometría. El caso de *El mejor amigo, el muerto*. Talía. *Revista de estudios teatrales*, *3*, 59-69. https://doi.org/10.5209/tret.74021

Cancelliere, E. (2000). Calderón e il Teatro di Corte. En M. L. Tobar (Ed.), *Atti della Tavola Rotonda sulla Singolarità Storica e Estetica di "La púrpura de la rosa" di Calderón de la Barca* (pp. 55-76). Armando Siciliano Editore.

Castro de Moux, M. E. (2001). Alquimia y gnosticismo en Fortunas de Andrómeda y Perseo de Calderón. En C. Strosetzki (Ed.), *Actas del V Congreso Internacional de la Asociación Internacio-*

[24] Está en marcha una nueva edición crítica del conjunto de *comedias* bajo la dirección de Ignacio Arellano dentro de la serie *Biblioteca Aurea Hispánica* de la editorial Vervuert. Actualmente, sin embargo, solo se han publicado 21 títulos. Este proyecto de edición puede considerarse la base textual más fiable; los principios de edición se aclaran en Arellano (2007). Además, las *Partes de las comedias*, aparecidas en vida de Calderón, están disponibles en una edición moderna en seis volúmenes a través de la editorial madrileña Fundación José Antonio de Castro, recién editada bajo la dirección de Luis Iglesias Feijoo (Calderón de la Barca, 2006-2010).

nal Siglo de Oro (AISO), Münster, 20-24 de julio de 1999 (pp. 319-330). Vervuert Verlagsgesellschaft.

Checa, J. (2010). *Pedro Calderón de la Barca: Amar después de la muerte. Edición y estudio*. Reichenberger.

Coenen, E. (2016). "La selva confusa" y "Cómo se comunican dos estrellas contrarias": comedias gemelas. *Revista de filología española*, *96*(1), 61-80. https://doi.org/10.3989/rfe.2016.03

Couderc, C. (2012). *Le théâtre tragique au Siècle d'or. Cristóbal de Virués, Lope de Vega, Calderón de la Barca*. Presses Universitaires de France.

Cuéllar, Á. (2022). Stylometry and Spanish Golden Age Theatre: An Evaluation of Authorship Attribution in a Control Group of Undisputed Plays. En C. Schöch, J. Calvo Tello, U. Henny-Krahmer, R. Hesselbach, & D. Schlör (Eds.), *Digital Stylistics in Romance Studies and Beyond*. [In press]

Ehrlicher, H. (2012). *Einführung in die spanische Literatur und Kultur des Siglo de Oro*. Erich Schmidt Verlag.

Ehrlicher, H., Lehmann, J., Reiter, N., & Willand, M. (2020). La poética dramática desde una perspectiva cuantitativa: la obra de Calderón de la Barca. *Revista de Humanidades Digitales*, *5*, 1-25. https://doi.org/10.5944/rhd.vol.5.2020.27716

Escudero Baztán, J. M. (2021). *Amor, honor y poder o el universo dramático de Calderón*. Iberoamericana Editorial Vervuert.

Greer, M. R. (1988). The Play of Power: Calderón's "Fieras afemina amor" and "La estatua de Prometeo". *Hispanic Review*, *56*(3), 319-341.

Jockers, J. (2013). *Macroanalysis. Digital Methods & Literary History*. University of Illinois Press.

Kroll, S. (2022). *Sonido y afecto en Calderón. Un estudio de las asonancias*. Reichenberger.

Lehmann, J. (2022). *Classification of Tragedies and Comedies in Calderón de la Barca's Comedias Nuevas* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7311502

Lowe, W. (2001). Towards a Theory of Semantic Space. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *23*, 576-581. https://escholarship.org/uc/item/0wk159m0

Maestro, J. G. (2003). Los límites de una interpretación trágica y contemporánea del teatro calderoniano: El príncipe constante. En M. Tietz (Ed.), *Teatro calderoniano sobre el tablado: Calderón y su puesta en escena a través de los siglos* (pp. 285-327). Franz Steiner Verlag.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and Their Compositionality. En C. Burges, L. Bottou, M. Welling, Z. Ghahramani & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 26: 27th Annual Conference on Neural Information Processing Systems* (NeurIPS: 26, Lake Tahoe, NV, 05-10.12.2013) (pp. 3111-3119). https://papers.nips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html

Parker, A. A. (1962). Towards a Definition of Calderonian Tragedy. *Bulletin of Hispanic Studies*, *39*,

222-237.

Parker, A. A. (1988). *The mind and art of Calderón. Essays on the Comedias*. Cambridge University Press.

Peirsman, Y., Geeraerts, D., & Speelman, D. (2010). The automatic identification of lexical variation between language varieties. *Natural Language Engineering*, 16, 469-491. https://doi.org/10.1017/S1351324910000161

Peña-Pimentel, M. A. (2011). *El Gracioso en el Teatro de Calderón. Un Análisis desde las Humanidades Digitales*. [Electronic Thesis and Dissertation Repository, 307] https://ir.lib.uwo.ca/etd/307

Peña-Pimentel, M. A. (2012). Aplicación de mapas de tópicos al análisis semántico de algunas comedias de Calderón. En J. L. Suárez (Ed)., *Calderón virtual. Anuario calderoniano* (pp. 115-130). Iberoamericana Editorial Vervuert.

Papay, S., Padó, S., & Thang Vu, N. (2018), Addressing Low-Resource Scenarios with Character-aware Embeddings. En Association for Computational Linguistics (Ed.), *Proceedings of the Second Workshop on Subword/Character Level Models* (pp. 32-37). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/W18-1204

Peirsman, Y., Geeraerts D., & Speelman, D. (2010). The automatic identification of lexical variation between language varieties. *Natural Language Engineering*, *16*(4), 469-491. https://doi.org/10.1017/S1351324910000161

Rosa, J. de la, Soto-Corominas, A., & Suárez, J. L. (2013). The Role of Emotions in the Characters of Pedro Calderón de la Barca's autos sacramentales. En: L. Beaven, & A. Ndalianis, *Emotion and the Seduction of the Senses, Baroque to Neo-Baroque*. (Conference, Melbourne, 27-29.11.2013; Studies in medieval and early modern culture, 59) (pp. 99-125). University of Melbourne.

Schöch, C. (2013). Fine-Tuning our Stylometric Tools. Investigating Authorship and Genre in French Classical Drama. En European Association for Digital Humanities (Ed.), *Digital Humanities Conference 2013* (DH 2013, Lincoln, Nebraska, 16-19.07.2013). European Association for Digital Humanities.

Schöch, C. (2017). Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama. *Digital Humanities Quarterly*, *11*(2), 1-53. http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html

Sullivan, H. W. (2017). *Calderón in deutschen und niederen Landen. Eine dreihundertjährige Rezeptionsgeschichte*. Matthes & Seitz.

Sullivan, H. W. (2018). *Tragic Drama in the Golden Age of Spain*. Reichenberger.

Tobar, M. L. (Ed.) (2000). *Atti della Tavola Rotonda sulla Singolarità Storica e Estetica di "La púrpura de la rosa" di Calderón de la Barca*. Armando Siciliano Editore.

Turney, P. D., & Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, *37*, 141-188. https://doi.org/10.1613/jair.2934

Valbuena Prat, Á. (1950). *Historia de la literatura española*. Vol 2: Los Siglos de oro (pp. 479-571).

Gustavo Gili.

Vega, F. L. de (1621). *Arte nuevo de hacer comedias en este tiempo. Dirigido a la Academia de Madrid*. Alonso Martín. https://books.google.de/books?id=lhh5ol6l4TsC (Obra original publicada 1609)

Ward, J. H. (1963) Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, *58*, 236-244.

Willand, M., & Reiter, N. (2017). Geschlecht und Gattung. Digitale Analysen von Kleists 'Familie Schroffenstein'. En A. Allerkamp, G. Blamberer, I. Breuer, B. Gribnitz, H. L. Lund, & M. Roussel (Eds.), *Kleist-Jahrbuch 2017* (pp. 177-195). J. B. Metzler.