# Zeitschrift für digitale Geisteswissenschaften

Autor*in:
Jörg Lehmann

Kontakt: joerg.lehmann@sbb.spk-berlin.de
Institution: Eberhard Karls Universität Tübingen
GND: 1054732310 ORCID: 0000-0003-1334-9693

Autor*in:
Sebastian Padó

Kontakt: pado@ims.uni-stuttgart.de
Institution: Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung
GND: 1033924393 ORCID: 0000-0002-7529-6825

Jörg Lehmann, Sebastian Padó

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

## Abstracts

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. 15 each 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 show stronger lexical cohesion. 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*.

In dieser Studie klassifizieren wir Komödien und Tragödien in einem Korpus von 112 Dramen Calderón de la Barcas, wobei wir Verfahren der distributionellen Semantik anwenden. Je 15 dieser *comedias nuevas* sind bereits von qualitativen Forscher*innen als Tragödien bzw. Komödien klassifiziert worden; bei weiteren 82 Dramen war die Klassifikation unbekannt. Es werden vier unüberwachte *document embedding*-Verfahren eingesetzt, die sich durch Matrixerstellung und -reduktion sowie durch die Berechnung von Ähnlichkeits- oder Distanzmatrizen voneinander unterscheiden. Die besten Ergebnisse – gemessen gegenüber der vorab vorgenommenen Klassifikation dieser Dramen – erzielt dabei jenes Klassifikationsverfahren, bei dem die stärkste Matrixreduktion vorgenommen wurde. Darüber hinaus wird eine kontrastive Vokabularanalyse mit *word embeddings* durchgeführt. Diese basiert entweder auf den Wortlisten der vier erprobten Verfahren oder auf der *Log-Likelihood*-Wahrscheinlichkeitsverteilung für zwei Subkorpora, die ausschließlich als Komödien oder Tragödien bestimmte Dramen enthielten. Dieser Arbeitsschritt ermöglicht die Identifikation von je 130 Begriffen, die für Komödien oder Tragödien diskriminativ sind. Das Ergebnis zeigt, dass die explorierten Verfahren Tragödien mit größerer Treffsicherheit identifizieren als Komödien, was darauf hindeutet, dass Tragödien mehr distinktive Merkmale aufweisen. Es zeigt sich aber auch, dass es angemessener ist, Klassifikationen wie ›Tragödie‹ und ›Komödie‹ als Pole zu denken, zwischen denen graduelle Unterschiede bestehen und in deren Übergangsbereich *comedias nuevas* enthalten sind, die in der Forschung als *tragicomedias* oder *comedias mitológicas* bezeichnet wurden.

# 1. Preface

Pedro Calderón de la Barca (1600–1681) counts, along with Félix Lope de Vega Carpio (1562–1635), as one of the most important playwrights of the Spanish baroque, also known as the ›Golden Age‹ (*siglo de oro*). His extraordinary and immense productivity rested on his unusual dexterity in re-using trivial plots, vocabulary symptomatic of any genre, and character types. However, Calderón's uniqueness arises from his ability to juxtapose the God-given balance of power with civil law and morality, and to present this opposition as an irresolvable, tragic conflict or an entertaining comedy. His works include 84 Corpus Christi plays (*autos sacramentales*), 112 *comedias* and 41 short pieces (*bailes, entreméses, jácaras, mojigangas*; contemporary terms also used by Calderón himself). A nearly complete collection of his works first appeared in the early 20th century from the Madrid-based publisher Aguilar.[1] Those of his *comedias* which had been published during his lifetime specified the dramas with terms such as *gran comedia* or *comedia famosa*. However, these descriptions did not differentiate between comedies and tragedies. This was in keeping with the use of language during the Golden Age, as the term ›comedia‹ was interchangeable with ›play‹ or ›theater piece‹: »Though the etymology of comedia is simple enough – a play of high spirits and laughter with a happy ending, – in Early Modern Spain the term comedia meant ›a play‹ or ›work for the stage‹ in a quite neutral sense.«[2] Because Calderón had never written any poetics himself, Lope de Vega's programmatical work *Arte nuevo de hacer comedias en este tiempo*[3] from 1609 is considered to be a contemporary reference by whose pragmatic rules Calderón generally oriented himself, despite some slight modifications. Here, Lope de Vega defines the *comedia nueva* as a play in three acts, and distinguishes the comedy as a fictional drama involving everyday people, from the tragedy as pertaining to members of the royal family or people of high descendance and being based on historic events. Furthermore, Lope characterizes the *comedia nueva* as a mixture of comedic and tragic

[1] Calderón de la Barca 1951–1956. This publication, however, does not conform to the standards of a historico-critical edition.
[2] Sullivan 2018, p. 33.
[3] Lope de Vega 1621.

elements, thus referring to the combination of both dramatic genres.[4] Thus, the Spanish playwrights of the 17th century had at their disposal a central poetological reference, which – superseding Aristotelian poetics – defined the ›Spanish style‹ as an original idea applying not only to comedy, but also to tragedy.

After a phase of degradation as being ›irregular‹ according to the doctrines of French classicism, the historical reception of the Spanish *comedia nueva* – and especially its understanding of tragedies – became vitally influenced through the German Enlightenment, the Romantic period and Idealism. Gotthold Ephraim Lessing (1729–1781) was one of the first in the German-speaking regions to recognize Calderón's work. He focused intensely on the tragedies of the Spanish Golden Age and implemented his theoretical aspirations on a practical level in a newly founded genre of the middle-class tragic drama. He was later followed by the Romantics Ludwig Tieck, August Wilhelm and Friedrich Schlegel, the brothers Grimm and Alexander and Wilhelm von Humboldt, who had all studied Spanish in Göttingen.[5] August Wilhelm Schlegel translated five of Calderón's plays for his *Spanisches Theater* (Vol. I: 1803, Vol. II: 1809) and examined Calderón in great detail in his *Vorlesungen über dramatische Kunst und Literatur* (Lectures on Dramatic Arts and Literature) in Vienna (1809). Wilhelm Joseph Schelling developed his own theory of tragedies in his presentation *Abhandlung über die Tragödie* (Essay on Tragedy) based on Calderón's work. Even Hegel and Schopenhauer grappled with the subject of Calderón, and thus it is no wonder that Walter Benjamin keeps returning to Calderón and his notion of the tragedy again and again in his *Ursprung des deutschen Trauerspiels* (Origin of the German Tragedy).[6]

While the interest in the German-speaking regions lay mostly on Calderón's tragedies and was, therefore, focused on only a few plays, it was first in the mid-20th century when serious attempts were made at examining and classifying the entire body of Calderónian *comedias nuevas.* It was initially the publishers of Calderón's *Obras completas*, who, in 1951, undertook a binary division of these theater pieces into *dramas* and *comedies*, thereby distinguishing between ›serious‹ relative to those resembling tragedies and ›light‹ relative to entertainment-oriented dramas. In this manner, the modern-day editors of the Aguilar publishing house quite obviously approached the provided examples of Calderón's *comedias* according to the poetic traditions of Antiquity, which, since the time of Aristotle, have been based on the clear separation of comedy and tragedy; however the editors proceeded with insufficiently explicit criteria.[7] At the same time, they posed a pivotal question with this differentiation, which has been heatedly discussed with opposing positions in the literary research of Calderon's work from the second half of the 20th century to the present day. The British Calderón school (Alexander A. Parker, Bruce Wardropper, Anthony Irving Watson, Henry W. Sullivan among others) was intensely occupied with Calderónian tragedies. Their attempts at classification were subjected to a rigorously methodical critique at the beginning of this millennium by the Spanish researcher Jésus G. Maestro, who commented, not without sarcasm, on the ›impotence of literary theory‹ regarding the dramatic genres and the ever-changing attributions accompanying them.[8] Now it was left to the British researcher Henry W. Sullivan to identify, from a qualitative perspective, twelve criteria according to which the tragic drama of the *siglo de oro* can be characterized. In doing so, Sullivan focused mainly on thematic traits (father-son conflicts, revenge and honor-based dramas), extra-literary indications (persons of high social standing),[9] characteristics of the plot (unfair judgements or death of the protagonist), or attributes of reception (creation of *eleos* and *pathos* or cathartic endings). He also formulated exclusionary criteria like the prevalence of themes such as redemption and damnation, and he also excluded martyr dramas, thus defining tragedies narrowly.[10] Within the framework of these criteria, Sullivan was able to identify at least 14 tragedies in the complete works of Calderónian *comedias nuevas.*

In light of the monumental works of Calderón it is, on the one hand, not surprising that the classification of the *comedias nuevas* – aside from the Aguilar edition – was never carried out comprehensively:[11] Which researcher is prepared to study and classify 112 dramas? At the same time, it is evident that just this sort of written work is suitable for the implementation of computational procedures. On the other hand, it must be understood that a data-based, computational classification of the entire body of the *comedias* has been rendered impossible until spring 2022, when all of them were made available in an electronic form.[12] Hence, Calderón's works – with the exception of only a few studies – have also not yet been analyzed with any methods provided by the

*digital humanities*, although such a massive corpus quite obviously lends itself to the examination of structural similarities among works in a particular genre or differences between dramas of varying genres.[13] Calderón's work stands out as a rare case in that such a large body of theater pieces was written by one author within a relatively short period during the 17th century.

The study at hand[14] represents an attempt, based on at least 112 *comedias,* to critically assess the validity of the distinction between the comedy and the tragedy among these dramas. This goes hand in hand with assessing the methodical possibilities made available by the digital humanities' application of *distributional semantics* procedures for this problem.[15] Because thus far only a small portion of the Calderónian *comedias* have been studied, and the majority of them remain entirely unexplored, we expect that the proven methods can deliver important indications for the classification of the plays which have yet to be thoroughly analyzed.

# 2. Methodology

## 2.1 Methodical Basis

Nowadays, the concept of distributional semantics is used widely in the realm of computational linguistics. The basic assumption is that the meaning of a word is established according to how much it is used and how often it co-occurs with other words within a specific context. Words and documents are represented in a high-dimensional space; semantic relationships are inferred from the similarities within that space. For the representation of documents, the frequencies (absolute or relative) of the words in each document are stored as matrices of vectors where each word corresponds to a column of the matrix and every document to a row. The cells of the matrix contain co-occurrence frequencies; pure frequencies are often replaced through degrees of statistical association, such as *pointwise mutual information* or *tf-idf* (*term frequency–inverse document frequency*), in order to counteract the Zipf distribution of words.[16] To represent the meanings of words, the same kind of matrix is created, with the target terms forming rows and contextual words forming columns. Such matrices can serve to compute the distances between single words or texts, to compare them to each other, to cluster them into groups, and to visualize them. As a rule, these very large matrices contain thousands of columns and are sparse, i. e. most of their elements are zero. This calls for reduction to a much smaller number of dimensions in order to be appropriate for the computation of distance or similarity matrices. The resulting low dimensional vectors are often referred to as *word* or *document embeddings* and are probably the most common practice for semantic representation in natural language processing (NLP). They are related to, but not identical to topic models. The reduction of dimensions is a purely technical requirement and hardly alters the underlying intention.[17]

The choice of a distributional approach for the task at hand is based on our starting assumption, namely the hypothesis that comedies and tragedies – in accordance with the treatment of each of the different themes – can be differentiated by observing word choice and word usage. Simply put, it can be expected that in Calderónian tragedies, terms such as ›honor‹, ›power‹ and ›death‹ strongly co-occur, while the comedies tend to combine words like ›love‹, ›disguise‹ and ›jealousy‹. This is quite obviously an approach that represents an oversimplification – narrative patterns or plot structures, however, cannot be characterized in this manner. At the same time, the wide success of approaches based on frequency and co-occurrence of words and common methods for author recognition demonstrates that such analyses allow for surprisingly deep understandings even of literary texts.

## 2.2 Data Basis

Beginning with the fourteen tragedies identified by Sullivan, yet another was added to the examined texts, which had apparently remained unknown to him: *Saber del bien y del mal.* [18] 15 further dramas, which were identified by qualitative research as comedies and which are often called *comedias cómicas* (or *urbanas* or *palatinas*),[19] make up the counterpart to the tragedies in this

---

[13] For example, Peña-Pimentel 2011; Peña-Pimentel 2012; de la Rosa et al. 2018; Ehrlicher et al. 2020.
[14] This study arose as a part of the project *QUOTE. Comprehensive Modeling of Conversational Contributions in Prose Texts*, sponsored by the German Research Community (Deutsche Forschungsgemeinschaft, project No. 350397899). The authors thank Prof. Dr. Hanno Ehrlicher (University of Tübingen), who commented on the first version of the article.
[15] Comparable studies on classical French drama have been thus far presented by, for instance, Schöch 2017 and Schöch 2013, who approached the subject with *topic modeling* and stylometric methods. For stylometric analysis of dramas in the *siglo de oro* cf. in particular Campión Larumbel / Cuéllar 2021 and Cuéllar 2022.
[16] Cf. Lowe 2001 for details.
[17] A short introduction is given in Jockers 2013, pp. 63–67.
[18] Cf. recently to this identification Escudero Baztán 2021, p. 21.
[19] See for the most recent overview of this classification Kroll 2022, pp. 63–65. Cf. also Calderón de la Barca 1951; Escudero Baztán 2021; Ehrlicher 2012; Maestro 2003; Parker 1988; Peña-Pimentel 2011; Tobar 2000; Valbuena Prat 1950.

body of work. The other 82 Calderónian *comedias* are available as full digital texts in modernized and normalized Spanish.[20] The spoken texts of the *dramatis personae* were extracted from all 112 plays and collected for analysis; stage instructions or similar additional texts were not included. The 15 tragedies were each marked with a T and a consecutive number, the comedies with a C, and the remaining 82 plays were marked »Test« and also numbered.[21]

## 2.3 Research Goal

In the absence of suitably large bodies of dramatic works beyond the Spanish-language world, the classification of genre with word or document embeddings is still relatively new.[22] Thus, the goal of our study is to explore various methods and combinations thereof, and to compare the results. We will compare four approaches, which all follow the same general unobserved schemes: 1) pre-filtering of the vocabulary; 2) calculation of document embeddings, and, if applicable, dimension reduction; 3) clustering of embeddings; 4) visualization und evaluation. Our corpus provides us with an excellent basis, as the categories are known in about a quarter of the plays, but not in the remaining dramas. In this manner, we can simultaneously review the quality of the process (on the basis of the known categories) and obtain findings on the yet unclassified dramas. We find this type of methodical comparison to be important, because it is known that the findings from unobserved distributional methods depend heavily on the parametrization of the process.[23]

## 2.4 Practical Application

All analyses were implemented with the statistics software R. The pre-processing of the texts was mostly carried out using the R package quanteda, as it also enables the exclusion of Spanish stop words, punctuation and numbers, and the conversion of the prepared corpus of texts to be processed in other packages. As was revealed in the course of exploration, only a small number (viz., 308) of Spanish stop words were retained in the quanteda package. One exploration showed that the exclusion of function words from the matrices did not lead to significantly different results, thus the stop word list was considerably expanded manually.[24] Furthermore, the analysis of the different methods employed, in particular the tf-idf statistics, showed that the grouping results were quite negatively affected by names of characters, places, and countries within the texts, also in their adjectivized form, as these elements of speech tend to reflect idiosyncrasies of single pieces rather than stereotypical genre characteristics. These proper names were likewise – primarily through the list of dramatis personae – compiled and removed from the texts; the number of terms to be excluded from the corpus thus rose above 800, additionally to the 308 stop words contained in the quanteda package. As a rule, the frequency of the words in each drama was calculated, subsequently the frequencies were normalized per document. This took place wherever the distance and similarity matrices for grouping were generated. When calculating the similarity between documents using cosine similarity this could be omitted, because they remain constant in relation to the vector lengths. Consistently throughout the analyses, work was done with inflected or conjugated forms of words; a lemmatization or a stemming of these words was not carried out. In this way linguistic information that might help in the classification of literary genres (and with respect to style, authorial signals or diachronic positionality) was preserved.

# 3. Results

## 3.1 Experiment 0

In a first exploration, we applied a well-established method, Skip-gram,[25] to the body of text in order to assess whether *word embeddings* could tell us something interesting about the text and which word pairs within the entire body of 112 dramas exhibited the highest number of similarities. We reduced the matrix to the 1,000 terms with the highest log-likelihoods and calculated the cosine similarity between all pairs of vectors. Cosine similarity, or more precisely, the cosine of the angle between two vectors, is a widely used measure of similarity which determines to what extent two vectors ›point‹ in the same direction in the high dimensional space. Cosine ranges between 0 and 1, and a high cosine indicates that two terms are found in similar contexts.

---

[20] For the most part, these dramas are available under the portal: Cervantes Virtual and the Association for Hispanic Classical Theater. A current overview of all sources can be found at: Estilometría aplicada al Teatro del Siglo de Oro. Because diacritical symbols used in modern Spanish can be used according to context, the spelling of certain terms may vary (ex.: solo / sólo – solo as an adjective means ›sole‹ or ›alone‹, whereas sólo as an adverb means ›barely‹ or ›merely‹).

[21] See the appendix below in which this abbreviation was removed and the results of the applied methods are presented.

[22] One exception is the study by Willand / Reiter 2017, cf. here pp. 190–194.

[23] Turney / Pantel 2010; Bullinaria / Levy 2007.

[24] These word lists are documented in the R code, which was published together with the body of dramas on Zenodo. Cf. Lehmann 2022.

[25] Mikolov et al. 2013.

Word pairings with a very high cosine similarity value of more than 0.75 are, for instance, »cielo« and »muerte« (heaven, death), »esperanza« and »desdichas« (hope, despair), »poder« and »temor« (power, fear), »poder« and »gusto« (power, taste), »honor« and »alma« (honor, soul) or »alma« and »muerte« (soul, death). One of the highest cosine similarity values, at 0.96, showed that the word pairing »honor« and »muerte« – honor and death – can be determined as a major theme throughout the entire body of work. Indeed, these first results proved to be surprisingly clear, in that, by using the Skip-gram algorithm, central themes in the Calderónian *comedias* could be identified, even when they deal with the intersection of social conventions (honor) and individuality (taste, soul, fear, social or actual death).

Conversely, word pairings like »honor« and »poder« (honor, power; 0.58), »amores« and »agravios« (love, infidelity, each in plural form; 0.69), »gracia« and »corte« (grace, court; 0.63) or »gracia« and »culpa« (grace, guilt; 0.60) showed lesser cosine similarity values. Cosine similarity values under 0.5 exhibit only weakly developed commonalities in the contexts; this could be observed for the word pairings »amar« and »honra« (loving, reputation), »muere« and »sepulcro« (he / she / it dies, grave), »muerte« and »engaño« (death, deceit), »mueran« and »suerte« (they may die, fate), »amores« and »honra« (love, reputation) and also »mentira« and »gracia« (lie, grace). First and foremost, it is apparent that the central themes in Calderón's works (»Amor, honor y poder«[26] – love, honor, and power) do not necessarily have to be interconnected with one another. This can be attributed to the fact that comedies and tragedies can be distinguished from each other through differing combinations of these terms. It is to be expected that the combination »honor« and »poder« is more characteristic of tragedies, and the combination »amar« and »honra« is more characteristic for comedies, but not for the entire body of work. We will come back to this point later.

## 3.2 Experiment 1

With the first experiment, our goal was to be able to explore the validity of the *document embeddings*. We take advantage of the known (or: labeled) tragedies and comedies to evaluate our document clusterings as follows, in the spirit of cluster purity[27] analysis: we assign each cluster to the class that the majority of documents with known affiliation belongs to. We then consider the other known classes of documents in this cluster, and compute purity, that is, the degree of agreement between these classes and the majority class, as a measure of success of our clustering. Our setup has the additional aspect that our data set includes documents for which the ›true‹ class is unknown. Since purity only considers documents with known classes, this makes the measure hard to interpret for clusters that consist predominantly or entirely of such documents. For such clusters — which we call underdetermined — we refrain from discussing purity in detail. After carrying out the preprocessing steps described above, we explored the following four methods: 1) Reduction of the matrix through the deletion of words according to their frequency and appearance within the texts; calculating the distance matrix according to relative frequencies, clustering with the Ward.D2 algorithm[28] based on the Euclidian distance. 2) Reduction of the matrix through the deletion of *sparse terms* which only appear in a few documents, calculation of the distance matrix based on relative frequencies, clustering based on the Euclidian distance with the Ward.D2 distance algorithm. 3) Part-of-speech tagging in each of the dramas, extraction of verbs, nouns and adjectives, calculation of the cosine similarity values between the documents, calculation of the distance matrix, clustering with the Ward.D2 distance algorithm. 4) Calculation of the tf-idf statistics, calculation of the cosine similarity values between the documents, calculation of the distance matrix and clustering with the Ward.D2 distance algorithm. We discuss the results of each method.

The first method represented a conservative approach: only the 1,094 words with a frequency > 120 and appearing in at least half of the documents were included. The document word matrix was filled with mere frequencies; no dimension reduction was carried out. The grouping was carried out through a clustering with the Ward.D2 distance algorithm. Figure 1 shows the resulting dendrogram. Recall that among the documents that form the leaf nodes of the dendrogram, some are known as comedies (CXX), some as tragedies (TXX), but most are unknown regarding their status (»Test«).

[26] Cf. Escudero Baztán 2021.
[27] Manning et al. 2008.
[28] Ward 1963.

Fig. 1: Ward.D2 clustering of 112 Calderónian Comedias. [Lehmann 2022]

Read from left to right, the first cluster represents a pure tragedy cluster which includes 29 dramas; 10 of these had already been characterized as tragedies. The third cluster from the left side depicts a pure comedy cluster; here 22 dramas are included, of which 10 had already been classified as tragedies. The two additional clusters must be described either as undefined or mixed clusters, as they either contain only 1 comedy (second cluster from the left, comprising 39 dramas) and therefore cannot be described as pure, or 4 comedies and 5 tragedies (the cluster to the right, comprising 22 dramas). Together, these two clusters contain more than half of the plays, namely 61 works. We conclude that with regard to the main research question, this approach does not appear to be especially effective, as only 20 of the 30 previously marked dramas (or 67%) were assigned in a clear fashion, while the remaining 10 comedies and tragedies mutually appeared in the clusters. However, the still relatively high dimensionality of the *document embeddings* makes a failure analysis challenging.

The goal of the second process is to create a low dimensional representation that is easier to interpret, in order to gain more insight into the distribution of the two genres. First, only terms which appear in at least 80% of all of the documents (i. e. in at least 90 plays) are retained; in other words, the sparsity is limited to 20%. This reduces the number of terms to a more compact total of 496. Again, a frequency-based word-document matrix is established and normalized, whereby the frequency of each of the remaining terms in each drama is divided by the sum of frequencies of *all* the words in the text. Finally, a distance matrix is established, based upon the Euclidian distance, and again, clustering is conducted using the Ward.D2 distance algorithm.
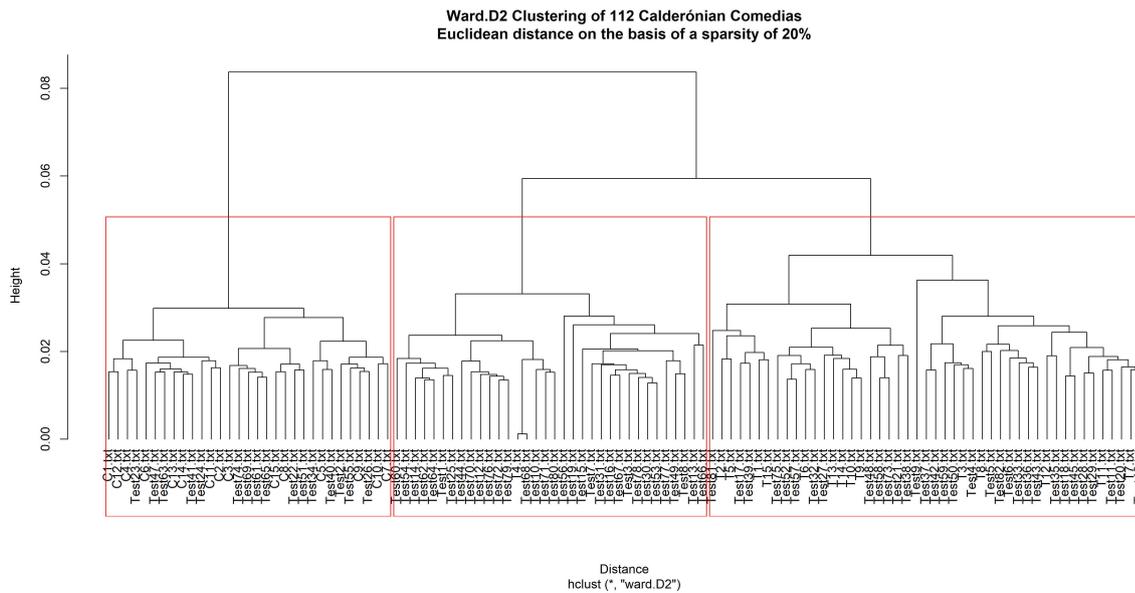
Fig. 2: Ward.D2 clustering of 112 Calderónian Comedias. Euclidian distance on the basis of a sparsity of 20%. [Lehmann 2022]

The dendrogram illustrates three clusters: In the first cluster to the left, all 15 comedies and 16 further dramas appear. The cluster on the right contains 14 tragedies and, likewise, 33 dramas of unknown classification. The cluster in the middle is mixed; it contains 1 tragedy (T4: *El mayor monstruo del mundo*) and 33 additional dramas of unknown classification. Through this process, which only deals with 496 words, 29 of 30 classified dramas, or 97%, were correctly assigned.[29]

Both of these automatic procedures, in which the fundamental matrices are reduced on the basis of word frequencies, establish a transitional zone between tragedy and comedy. This observation presents us with the question of whether it would be more appropriate, in light of distributional semantics, to consider classifications like ›tragedy‹ and ›comedy‹ as poles between which gradual differences appear, showing the resulting overlap in regard to the applied word selection. In the matter of Calderónian dramas, this seems quite sensible, as themes such as ›honor‹ and ›power‹ can just as well be included in comedic plots as in those of the famous honor tragedies.

Comedies may also present serious subjects in a lighthearted, entertaining manner. For example, power struggles between royal families can be indirectly alluded to within the framework of a mythological play; the allegory would have been quite understandable for the court audience at the time.[30]

One possible fundamental critique on simple *document embedding* methods, like those we have observed thus far, is the total absence of linguistic structure. For this reason, we made the decision to subject all of the dramas to *part-of-speech tagging ,* including only verbs, nouns and adjectives from each play in the corpus for clustering.[31] For testing the third procedure, therefore, a second corpus is established, in which each of the drama texts include only verbs, nouns and adjectives in their basic forms. All proper names are once more filtered out of the matrix created for this purpose – they had been falsely recognized as adjectives – and subsequently a calculation is made, based on the non-normalized frequencies of the cosine similarities. This similarity matrix is converted to a distance matrix and, once again, clustered with the Ward.D2 algorithm. The results are depicted in a dendrogram.

---

[29] Basically, we attempted to alter only one parameter between each of the analyses, thus using the Euclidian distance. As an alternative, during the second procedure, we also used the Manhattan distance, whereby the distance is defined by the sum of absolute values. The results were clearly less satisfactory than the above representations resulting from the use of the Euclidian distance: Only two thirds (67%) of all previously identified tragedies and comedies were correctly clustered.

[30] This possibility was already mentioned by Greer 1988 in an example from *Fieras afemina amor*.

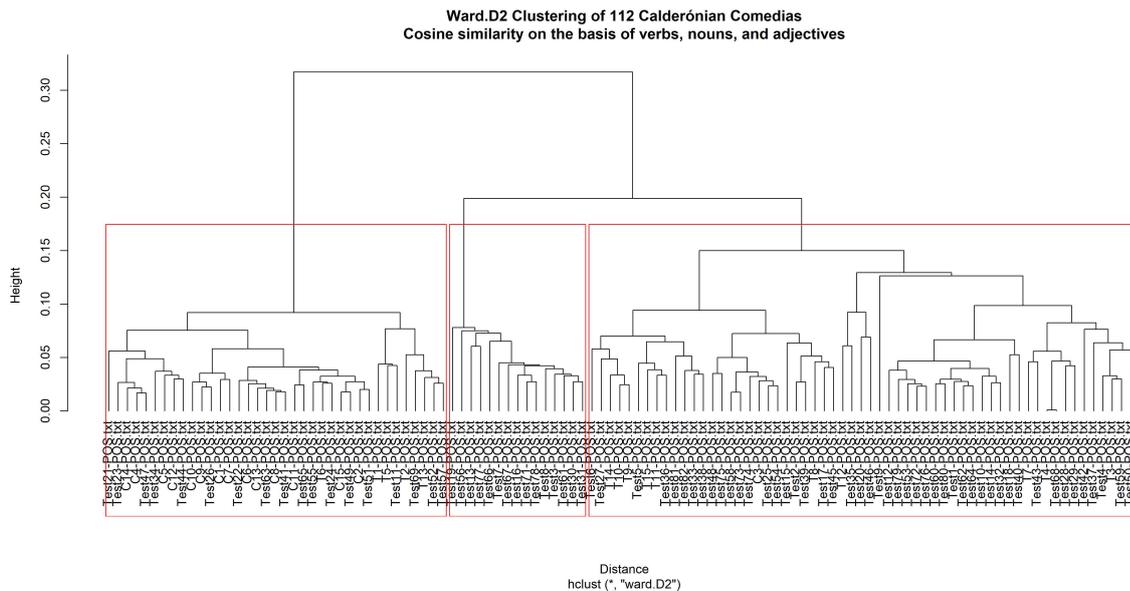[31] This kind of method was used by Willand / Reiter 2017, pp. 191f.

Fig. 3: Ward.D2 clustering of 112 Calderónian Comedias. Cosine similarity based on verbs, nouns and adjectives. [Lehmann 2022]

The first cluster to the left, which might be identified as a comedy cluster, contains 14 comedies, 5 tragedies (T1: *A secreto agravio, secreta venganza*; T2: *El alcalde de Zalamea*; T5: *El médico de su honra*;[32] T6: *El pintor de su deshonra*; T13: *Las tres justicias en una*) and 18 additional plays of unknown classification. The cluster to the right is mostly a tragedy cluster, because it contains 10 tragedies and 49 additional plays, but also 1 comedy (C3: *El encanto sin encanto*). In the middle between these two categories is an undefined cluster, containing 15 plays marked with »Test«. With regard to the plays identified thus far as tragedies and comedies, 80% of these dramas were correctly clustered; however, this result applies only if clusters are identified by the majority of previously identified dramas.[33]

Taking into consideration the previously tested methods, it seems advisable to focus on every term that carries meaning and thus leads to a differentiation between the categories. The fourth method we tried was based on the tf-idf statistics, thus underlying a measure of association commonly used in *text mining*, whereby terms can be evaluated for their significance within a document or body of work. With the tf-idf statistics, the weight of each term per document is calculated; the *term frequency* (*tf*) is multiplied by the *inverse document frequency* (*idf*). The latter depends not on individual documents, but rather on the total number of all documents in the corpus. In this way, the tf-idf statistics considers the relative significance of words which appear frequently in the corpus to determine how relevant the term is for a document within the corpus under study. Once more, the proper names are removed, the cosine similarity for the vectors is calculated, the similarity matrix is converted into a distance matrix and clustering is carried out with a Ward.D2 algorithm. The results are depicted in a dendrogram.

---

[32] This outcome is especially interesting, because, according to Couderc 2012, p. 104 *A secreto agravio, secreta venganza* and *El médico de su honra* can be described as tragicomedies and *A secreto agravio, secreta venganza* is the only play by Calderón which uses the term »tragicomedia« (tragicomedy) in the spoken text.

[33] As an alternative, a normalized matrix was established and a Ward.D2 clustering based on the Euclidian distance was carried out. The results are clearer, since 4 tragedies as well as 14 comedies were assigned to a non-mixed cluster. However, the remaining 11 tragedies and 1 comedy formed a mixed cluster, so that all in all only a purity of 60% in the clustering was reached. See the R code in Lehmann 2022.
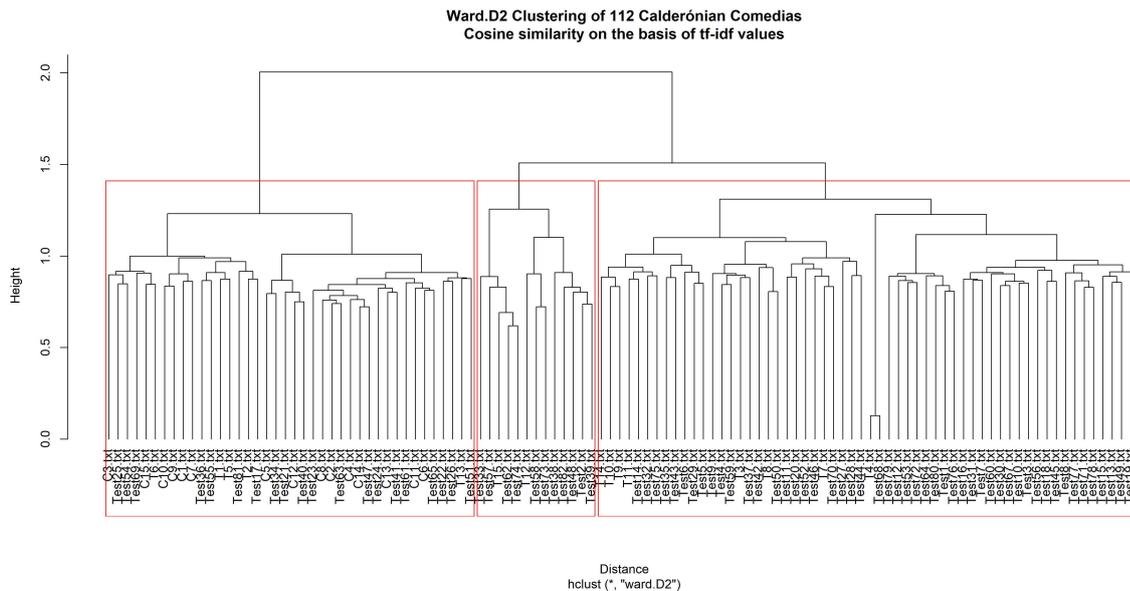
Fig. 4: Ward.D2 clustering of 112 Calderónian Comedias. Cosine similarity on the basis tf-idf values. [Lehmann 2022]

This image shows three clusters: The first one to the left can best be described as a comedy cluster. In addition to all of the 15 comedies, however, it also contains 5 tragedies, exactly the same five ones as in the part-of-speech based analysis conducted previously (T1, T2, T5, T6, T13), as well as 20 other dramas. The cluster on the right, with 8 tragedies and 51 further plays, can be considered a tragedy cluster. The smallest one in the middle is hard to define, since it is only weakly determined and thus cannot be understood as pure; it contains only 2 dramas clearly identified as tragedies and eleven others of unknown classification. In comparison with the dramas already identified as tragedies or comedies, this result shows that 8 of 15 tragedies and, respectively, all comedies have been clustered correctly; this correlates to a recognition rate of 76%.[34] Compared against the models considered above, this recognition rate seems to be satisfactory.

The four methods explored here differ by the choice of data as well as by the choice of distance or similarity metrics. Three of the four generated robust to very good results. The process of employing the strongest matrix reduction produced the best findings. However, only one approach yielded a clustering result that would arguably approximate the classification of researchers applying qualitative analyses.

## 3.3 Experiment 2

In a second experiment, we assess to what extent the document clusterings we found in the first experiment were based on word choice or word use being consistent with the two genres. To do so, we analyze the word lists upon which the clusters found by the four methods were based. In addition, we calculate the log-likelihood distribution over the vocabulary for the sets of (predicted) comedies and tragedies of each method. This approach determines the 200 words with the highest log-likelihood values for each genre, and these lists can be compared across methods (contrastive vocabulary analysis with *word embeddings*).

Recall that the first procedure in experiment 1 (Ward.D2 clustering based on the Euclidian distance between normalized word frequencies) created a clustering in which only the first and the fourth clusters could be clearly assessed as comedy or, relatively, tragedy clusters. For both of these clusters, the probability margin for each word is evaluated based on the previously established matrix, and the 15 terms with the highest probability margin for each were investigated. These 15 selected terms for both comedy and tragedy clusters with the highest probability margins give an impression of the cluster formation. For the comedy cluster, the terms »don«, »casa«, »calle«, »papel«, »caballero«, »puerta«, »dama«, »padre«, »hermano«, »saber«, »cuarto«, »amigo«, »hombre«, »sé«, and »señora« (esquire, house, street, paper, knight, door, lady, father, brother, knowledge, room, friend, man, I know, and madam) appeared. Very interesting is the word »papel«, since it points to the paper or card fanning the intrigue; however, beyond this term, the word list does not seem to be significantly distinctive of comedies. By contrast, for the

---

[34] Here, alternatively, a Ward.D2 clustering was also carried out based on the Euclidian distance. The result shows five clusters, three of which contain four dramas labeled with »Test«. The remaining two clusters consisted of one mixed cluster containing 15 comedies, 12 tragedies, and 40 further plays; and another cluster containing 3 tragedies as well as 38 further plays. These results confirm the unreliability of this approach with respect to clustering.

tragedy cluster, the words »rey«, »muerte«, »dios«, »cielo«, »hoy«, »vida«, »sol«, »valor«, »mar«, »tierra«, »gran«, »rigor«, »mundo«, »quiero«, and »poder« (king, death, God, heaven, today, life, sun, value / valor, sea, earth, grand, severity, world, I want, and power) were especially frequent. At any rate, people of high social standing, death, God, valor and power stand out as being characteristic terms relating to these storylines.

The 496 words selected for their sparsity of 20% enable a preview of terms which carry a strong distinction with regard to the separation of comedies and tragedies. For the comedy cluster, meaningful terms like »don«, »casa«, »dama«, »calle«, »puerta«, »sé«, »señor«, »caballero«, »bien«, »cuarto«, »papel«, »señora«, »saber«, »amigo«, and »celos« (esquire, house, lady, street, door, I know, lord, knight, good, room, paper, madam, knowledge, friend, and zeal) are present. For the tragedy cluster, words like »rey«, »señor«, »dios«, »hoy«, »muerte«, »cielo«, »sol«, »quiero«, »rigor«, »mundo«, »gran«, »valor«, »alma«, »viento«, and »sangre« (king, lord, God, today, death, heaven, sun, I want, severity, world, grand, value / valor, soul, wind, and blood) appear. At first glance, the high degree of consistency of both lists of words from the first and second procedures may come as a surprise. Then again, it appears that the high degree of purity in the clustering of the second procedure quite obviously depends upon the condensed and precise selection of distinct terms.

With regard to the third procedure – based upon a *part-of-speech* tagged corpus – the most frequent words found in the clusters in the underlying matrix illustrate why it does not lead to compelling results: Not surprisingly, the most frequent words here are the verbs »ser« (to be) and »haber« (to have), followed by a list of much less frequent additional verbs, like »ver«, »decir«, »estar«, »dar«, »poder«, »saber«, »hacer«, »tener«, »ir«, »querer«, »venir« (seeing, saying, being, giving, being able, knowing, doing, having, going, wanting, and coming). This is then followed by a list of nouns, like »señor«, »vida«, »cielo« or »don« (lord, life, heaven or esquire). In light of the fact that these frequently used words seem to have little ability to distinguish between comedies and tragedies, the results of the clustering can be described as rather poor.

In the fourth procedure – based on the tf-idf matrix – an approach analogous to methods 1 and 2 is applied. The 15 terms that show the highest probability margin within the comedy cluster are: »don«, »doña«, »tapada«, »hermana«, »calle«, »hermano«, »coche«, »amiga«, »anoche«, »papel«, »cuarto«, »aposento«, »reja«, and »casa« (esquire, lady, veil, sister, street, brother, carriage, friend, last night, paper, room, chamber, grid, and house). In the tragedy cluster, terms such as »arma«, »dioses«, »cristianos«, »templo«, »montes«, »cueva«, »ciencias«, »cruz«, »muro«, »reino«, »pastor«, »rey«, »cristiano«, »cajas«, and »guerra« (arms, gods, Christians, temple, mountains, cave, sciences, cross, wall, kingdom, pastor, king, Christian, crates, and war) are characteristic. While the frequent terms selected for the comedy cluster seem, for the most part, to be less discriminating, save for the typical allusions to veiling and masking or intrigue through forgery, the terms relating to tragedy reflect, at least, military and Christian themes as well as the aristocratic descent of the protagonists.

An open question at this point is how robust these methods are. Thus, in the next step, we test the word lists created in the steps above and base them on a larger body of works. We extend our data basis to clusters, expand the body of plays identified as comedies or tragedies and create two somewhat larger subgroups. From the dramas hitherto marked as »Test«, we choose 16 which were unanimously clustered as being ›tragedy‹ by all four procedures, as well as ten which were unanimously clustered as ›comedy‹. For the comedies, we corroborated this classification on the basis of secondary literature;[35] moreover, all of these dramas were included in the collection of comedies by the editors of the Aguilar edition. In this manner, we generate two new subgroups, one for tragedies, containing 31 plays, and one for comedies, containing 25 plays.[36] Both of these subgroups are converted into matrices using the prevalent preprocessing techniques, whereby all of the terms found in less than four of the plays are filtered out. For the remaining words, the 200 most informative for each subgroup are identified for inclusion, using the log-likelihood function, with which discriminative terms can be found. The comparison of the results for each subgroup shows that only 70 terms appear in both lists, while 130 terms for each (almost exactly two-thirds) are discriminative for either the tragedy or the comedy subgroup.

The analysis of these 130 discriminative terms for each subgroup proves to be very revealing. In the case of the comedies, we discover references to certain themes (»ama«, »amiga«, »carta«, »celoso«, »desdichas«, »desengaño«, »escondido«, »favor«, »joyas«, »juego«, »máscara«, »papeles«, »secreto«, »tapada«, »vestido« – mistress, girlfriend, letter, jealous, misfortune, disappointment, veiled, favor, jewelry, game, mask, papers, secret, hidden / stashed, disguise), typical indications relating to the

---

[35] Nearly all of these dramas fall in the category »Comedias cómicas« described by Kroll 2022, pp. 64–65. However, there are two exceptions: In contrast to Kroll's estimation, who puts *No hay cosa como callar* into the category »Tragedias y dramas de honor«, we classify this drama as comedy, since all the four employed methods were in agreement. By comparison, we dismissed *Las manos blancas no ofenden* from the list of comedies, since the estimation of Valbuena Prat 1950, who counts this play amongst »obras exclusivamente cómicas« (p. 541), was not corroborated by the procedures applied by us.
[36] Cf. for a comparative method Peirsman et al. 2010.

mythological background of the comedies (»astrólogo«, »duende«, »forastero«, »jardines«, »ninfas« – astrologer, elf / gnome, foreigner, gardens, nymphs) and also the appearance of some rather surprising terms (like »enemigo«, »pendencia«, »razón« or »saber« – enemy, brawl, reason or knowledge).

By contrast, among the tragedies we find references to the (mostly high) standing of the characters (»convento«, »corona«, »emperador«, »esclavo«, »infanta«, »infante«, »majestad«, »reina«, »reinar«, »reino«, »rey«, »tirano«, »villano« – cloister, crown, emperor, slave, infant, infanta, highness, queen, ruling, kingdom, king, tyrant or villain), the contents of the plot (»cristo«, »cruz«, »desdichado«, »divina«, esperanza, »gloria«, »laurel«, »lealtad«, »libertad«, »morir«, »poder«, »salud«, »sangre«, »traición«, »triste«, »triunfo«, »venganza«, »victoria« – Christ, cross, misery, divine, hope, fame, laurel, devotion, freedom, dying, power, health, blood, treason, sad, triumph or revenge, victory) and a few surprises as well (»ciencias«, »enamorado«, »sueño« – sciences, enamored or dream). Altogether, the word lists determined log-likelihoods in the two subgroups outline the contents of the comedies and tragedies much more precisely than the word lists based on each cluster.

# 3.4 Experiment 3

In our final experiment, we move beyond the analysis of documents in terms of words, as in experiment 2, to an analysis of the usage of individual words across the two genres. For this purpose, we used the embedding method fastText[37] and the R package of the same name. In each subgroup, the ten nearest neighbor terms of interest are established, so that each word which was identified as pertaining to both genres is visible, along with the terms found closest to it within the text. In contrast to Skip-gram, fastText is more appropriate for smaller bodies of text, as it does not compute an *embedding* for each word. Instead, *embeddings* for parts of words are calculated (for instance, for »honor«: »hon«, »ono«, »nor«, etc.) and accumulated to create an *embedding* for the whole word. In this way, more robust representations emerge for rarely used or unknown words.[38]

In order to contrast the terms in each subgroup, we will illustrate in the following the ten nearest neighbor terms per subgroup together with the similarities for each, whereby the maximum possible similarity is represented by the number 1.

The keyword »honor«, which is found not only in comedies, but also in tragedies, when assessed within the comedy subgroup, shows no common neighboring terms in the tragedy subgroup, nor were they found for the word »hado« (fate). In other words, both terms are used in comedies and tragedies, but within completely different contexts according to each. It becomes apparent that the terms ›honor‹ and ›fate‹ appearing in tragedies are more clearly outlined within the context and the meaning of the terms more precisely defined. For example, ›honor‹, within the context of the tragedy, refers to the loss thereof, or defamation, for which the remedy is obviously associated with possible death.

---
[37] Bojanowski et al. 2017.
[38] Papay et al. 2018.

| Comedia | Tragedia |
|---|---|
| honor | |
| - pundonor 0.81 (honorability)<br>- ofrecer 0.80 (offer)<br>- lograr 0.79 (achieve)<br>- honrar 0.79 (to honor)<br>- obedecer 0.78 (obey)<br>- menor 0.78 (minor)<br>- reconocer 0.78 (acknowledge)<br>- rencor 0.77 (grudge)<br>- confesar 0.77 (confess)<br>- ofender 0.77 (offend) | - satisfación 0.81 (satisfaction)<br>- sujeción 0.78 (subjection)<br>- oración 0.77 (prayer)<br>- rigor 0.76 (rigor)<br>- maldición 0.76 (curse)<br>- opinión 0.75 (opinión)<br>- satisfecha 0.75 (satisfied)<br>- satisfacción 0.75 (satisfaction)<br>- honra 0.75 (honor)<br>- acción 0.75 (action) |
| hado (fate) | |
| - hallado 0.92 (found)<br>- amado 0.91 (loved)<br>- hablado 0.91 (spoken)<br>- madrugado 0.90 (gotten up at dawn)<br>- echado 0.90 (thrown)<br>- mirado 0.89 (looked)<br>- negado 0.89 (denied)<br>- pecado 0.89 (sinned)<br>- tocado 0.87 (touched)<br>- enfadado 0.87 (angry) | - estimado 0.92 (estimated)<br>- librado 0.91 (liberated)<br>- engañado 0.90 (enchanted)<br>- sobrado 0.88 (surplus)<br>- nombrado 0.88 (named)<br>- tratado 0.88 (treated)<br>- rendido 0.87 (surrendered)<br>- desengañado 0.87 (disenchanted)<br>- mostrado 0.87 (shown)<br>- estrado 0.87 (stage) |

Tab. 1: 10 nearest neighbor terms for »honor« and »hado«. [Lehmann / Padó 2022]

The many similar word endings in this table may be baffling at first glance, but hardly surprising: All of Calderón's plays are written in verses. Through this metric alone, the selection of possible neighboring words is drastically limited.[39] To make things worse, the similar inflections and conjugations of the Spanish language also left Calderón with a very narrow selection of possible neighboring words when composing his dramatic works.

Other terms which were used in both subgroups also produce a similar pattern. The words »fineza«, »justicia«, and »amistad« (nicety, justice, friendship) yielded only one or two common neighboring words within both subgroups (represented in bold type); these terms are found in both comedies and tragedies alike, but within very different contexts. While these three terms within the comedic context tend to reflect the profane, their appearance in the tragic context reflects the formal authority of the court and its jurisdiction as well as seriousness and the realm of divine providence and justice.

---

[39] An example from the tragedy *La gran Cenobia* , where »honor« rhymes with »rigor«: »[Libio:] Por verme con alto honor, / La muerte á Abdenato di, / Mi misma sangre vendí, / A mi patria fui traidor. / Llegó el rigor / A castigarme, y á ser / Mi verdugo osado y fuerte; / Pues advierte, / ¿Qué tengo ya que perder, / Perdido el miedo á la muerte?« There are also examples of two words that rhyme within a verse, such as in the comedy *Cuál es mayor perfección, hermosura o discreción*, where »honor« rhymes with »pundonor«: »[Beatriz:] ¿Félix, restado su honor / y yo sabidora de ello / y no tratar de enmendarlo? / Eso no; que por mi mesmo / pundonor debo acudirle.«

| Comedia | Tragedia |
|---|---|
| **fineza (nicety)** | |
| - **firmeza** 0.84 (firmness) | - fiereza 0.84 (fierceness) |
| - fianza 0.81 (pledge) | - gloria 0.78 (glory) |
| - importuna 0.81 (important) | - peregrina 0.77 (pilgrim) |
| - fina 0.80 (fine) | - indignación 0.77 (indignation) |
| - impida 0.80 (impede) | - insignia 0.77 (insignia) |
| - implica 0.79 (implies) | - ofrecí 0.76 (offered) |
| - naturaleza 0.79 (nature) | - grandeza 0.76 (greatness) |
| - nobleza 0.78 (nobility) | - **firmeza** 0.75 (firmness) |
| - templanza 0.78 (temperance) | - imperial 0.75 (imperial) |
| - belleza 0.77 (beauty) | - ignorancia 0.75 (ignorance) |
| **justicia (justice)** | |
| - **justa** 0.83 (just) | - **justa** 0.83 (just) |
| - hidalga 0.78 (noble) | - justiciero 0.82 (avenging) |
| - acompañada 0.77 (accompanied) | - licencia 0.80 (licence) |
| - **malicia** 0.77 (malice) | - precia 0.79 (precious) |
| - salida 0.76 (departure) | - milicia 0.79 (militia) |
| - diligencia 0.75 (diligence) | - **malicia** 0.78 (malice) |
| - hidalguía 0.75 (nobility) | - usted 0.77 (you) |
| - historia 0.75 (history) | - gusta 0.77 (like) |
| - dispensación 0.75 (dispensation) | - estudiar 0.77 (study) |
| - traición 0.75 (treason) | - condición 0.76 (condition) |
| **amistad (friendship)** | |
| - dad 0.85 (giving) | - acudid 0.82 (attend) |
| - vanidad 0.83 (vanity) | - **calidad** 0.82 (quality) |
| - **mitad** 0.83 (half) | - ofrezca 0.81 (offer) |
| - debéis 0.83 (owe) | - seguridad 0.81 (safety) |
| - decid 0.81 (decide) | - fealdad 0.77 (ugliness) |
| - **calidad** 0.81 (quality) | - temeridad 0.77 (recklessness) |
| - mirad 0.80 (look) | - **mitad** 0.77 (half) |
| - libertad 0.80 (freedom) | - sacad 0.76 (pull) |
| - perdonad 0.79 (forgive) | - firmeza 0.76 (firmness) |
| - podáis 0.79 (can) | - salid 0.76 (get out) |

Tab. 2: 10 nearest neighbor terms for »fineza«, »justicia« and »amistad«. [Lehmann / Padó 2022]

However, other terms clearly show overlaps with regards to the nearest neighbor terms; for instance, »celos«, »gusto« or »muera« (zeal / jealousy, taste, he / she / it dies) each share three or four nearest neighbor terms within the ten words in the selection.

| Comedia | Tragedia |
|---|---|
| **celos (zeal, jealousy)** | |
| - **celosos** 0.91 (jealous) | - consuelos 0.91 (consolations) |
| - **recelos** 0.90 (suspicions) | - **recelos** 0.91 (suspicions) |
| - duelos 0.89 (duel) | - **celosos** 0.90 (jealous) |
| - **cielos** 0.85 (heavens) | - antojos 0.89 (cravings) |
| - puestos 0.84 (posts) | - pueblos 0.89 (villages) |
| - palos 0.83 (sticks) | - regalos 0.88 (gifts) |
| - dellos 0.83 (from them) | - demos 0.88 (we give) |
| - desconsuelos 0.82 (hopelessness) | - **cielos** 0.87 (heavens) |
| - opuestos 0.82 (opposites) | - caballos 0.87 (horses) |
| - laberintos 0.82 (mazes) | - verlos 0.87 (see them) |
| **gusto (taste)** | |
| - admito 0.87 (admitted) | - **justo** 0.87 (just) |
| - visto 0.86 (seen) | - desprecio 0.85 (contempt) |
| - susto 0.86 (scare) | - precio 0.84 (prize) |
| - **justo** 0.84 (just) | - justiciero 0.84 (righteousness) |
| - gasto 0.84 (expense) | - **disgusto** 0.83 (displeasure) |
| - **disgusto** 0.84 (disgust) | - precepto 0.82 (precept) |
| - pedido 0.83 (order) | - preciso 0.82 (precise) |
| - considero 0.82 (consider) | - profano 0.82 (profane) |
| - adentro 0.82 (in) | - favorecido 0.82 (favored) |
| - pecado 0.82 (sin) | - convencido 0.82 (convinced) |
| **muera (he / she / it dies)** | |
| - **muriera** 0.89 (dying) | - viviera 0.94 (living) |
| - **muerta** 0.89 (dead) | - **muriera** 0.94 (dying) |
| - defuera 0.85 (outside) | - **muerta** 0.92 (dead) |
| - **muralla** 0.85 (wall) | - **muralla** 0.91 (wall) |
| - muestra 0.84 (sample) | - diera 0.90 (giving) |
| - **manera** 0.83 (way) | - madera 0.90 (wood) |
| - mira 0.82 (look) | - **manera** 0.90 (way) |
| - enferma 0.81 (sick) | - viera 0.90 (watching) |
| - dondequiera 0.81 (anywhere) | - hermosura 0.89 (beauty) |
| - cólera 0.81 (anger) | - matara 0.89 (kill) |

Tab. 3: 10 nearest neighbor terms for »celos«, »gusto« and »muera«. [Lehmann / Padó 2022]

This analysis illustrates that the differences between tragedies and comedies do not merely consist of different vocabularies, but rather, that even shared vocabularies are substantially *used in a different way*. The more central for the genre, the more distinguishable the usage – at least, this is the tendency our results have shown so far.

# 4. Discussion of the Results and Outlook

The comparison of the methods shows that with two of them – clustering of dramas on the basis of verbs, nouns, and adjectives and clustering on the basis of tf-idf values – results can be reached that approximate expert judgments. Both methods are considered standard procedures in text mining. In order for the clustering to reach a purity of 70% and beyond, however, comprehensive filtering was needed, extending beyond the usual punctuation and stop words to further function words, proper nouns and their adjectivized forms. A part of the latter can only be manually assembled for each corpus under study, which requires considerable effort. A rather good purity of the clustering can be reached fairly fast by conducting a massive reduction of the output matrix to a sparsity of 20%, thus considering only terms which appear in at least 80% of all of the documents.

The preliminary observations of this study considering the comparison of the four explored methods permit us to identify further dramas of each category (sixteen tragedies and ten comedies) which could be regarded, with a high probability, as being either tragedies or comedies. They also point to characteristic mixtures of the vocabularies in use as well as to contradictory results. This particularly concerns comedic passages in the dramas – even when they appear within a tragedy – but also any terms that reflect themes that are typical for comedies or tragedies, extra-literary attributes or plot characteristics.

One particular example would be *Amor, honor y poder*, a title unknown to the authors in this study before the analysis began. Though it is commonly classified as a comedy because of its happy ending, the intrigue deals with unhappy relations between two pairs of characters and is therefore dominated by a semantics typical of tragedies. While the methods employed in this study all classify this drama as a tragedy, another exception is formed by *No hay cosa como callar*. Again, all the four procedures classify this drama unanimously, in this case as a comedy, and so does the Aguilar edition. The judgments of qualitative research, however, are more divided: While Alexander A. Parker classified it in 1962 as a tragedy, he later revised his judgment and described it as a »comedy of intrigue«, and Simon Kroll puts it into the section »Tragedias y dramas de honor«.[40] Certainly, the analysis conducted here will inspire further debates, since such variations in the classification of a drama may be resolved by a differentiated examination: The vocabulary in *No hay cosa como callar* may be one typical for comedies, but the plot as well as other qualitative criteria might support its classification as a tragedy.

Also interesting is the insight that the Calderónian tragedies, obviously because of the way the words are used within the text, are much more reliably identifiable than the comedies. This is underlined by the way in which all of the four applied methods identified the group of so-called *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, and Origen, pérdida y restauración de la Virgen del Sagrario*. All these dramas are consistently marked by the use of a tragic vocabulary. On the other hand and with regard to the comedies, it is quite obvious that they are much harder to define than tragedies. This is true, for example, with respect to a group of comedies which are frequently regarded as *comedias mitológicas*. The mythological plays *El castillo de Lindabridis, El mayor encanto amor, La puente de Mantible*, and *Los tres mayores prodigios* exhibit very strong tragedy signals in our analysis, whereas most other dramas classified as *comedias mitológicas* exhibit mixed signals.[41]

Certainly, with regard to dramas stipulated on the basis of our analysis which, up to now, have received very little attention, the binary separation of *dramas* and *comedias* previously conducted by the publishers of the Aguilar edition must be viewed with a critical eye. A good example for this is provided by *Amar después de la muerte*, which stood out through the use of tragic vocabulary as identified by the most precise clustering approach (method 2). This classification was verified by the historical-critical edition presented by Jorge Checa.[42] Since Checa, in the preface of his analysis, discusses a series of criteria regarding the designation of tragedies according to Parker and Sullivan, this insight presents an invitation to qualitatively working researchers to work systematically and to consistently implement these established criteria for classification on an entire sequence of plays. The status of the group of dramas called *comedias mitológicas* – as with those recognized by Parker and Sullivan as being »on the brink of tragedy«[43] – should therefore be discussed anew with regard to their designated categories and the vocabularies used. The same is true concerning the scarcely examined group of dramas which can be classified as »tragicomedias«. The intermediate area found between comedies and tragedies throughout these methods points to this in an emphatic way. In the sense of the digital humanities, this conclusion represents an invitation to qualitative researchers to take a deeper look at the texts they have already examined and to create lists of characteristic words for each category to be distinguished.

The approach performed through distributional semantics contributes only one factor among others – albeit an arguably important one – to the classification of plays, in particular when, as is the case here, lexical and semantic analyses go hand in hand. This is especially relevant in view of the large number of works which have yet been only scarcely researched or not at all. The systematic comparison of various methods, as carried out here, presents the opportunity to better evaluate the results of heterogeneous corpora (plays by various playwrights or from different centuries). The implementation of these tested procedures on, for example, all available dramas in the *siglo de oro,* would provide a broader basis for the achieved results upon which characteristic lexica for comedies and tragedies could be identified. Precisely, however, the example of Calderón with his

---

[40] Cf. Parker 1962, p. 228; Parker 1988, pp. 181–182; Kroll 2022, p. 63.
[41] For the assessments of these works as *comedias mitológicas*, see Kroll 2022; Castro de Moux 2001; Greer 1988; Cancelliere 2000; Arellano 2000; Peña-Pimentel 2011.
[42] Checa (Ed.) 2010.
[43] Cf. Parker 1988, pp. 58, 181, 182; Sullivan 2018, pp. 70, 316, 321.

112 *comedias nuevas* illustrates that the methods explored here provide qualitative researchers with information, which may stimulate further analyses. Potentially, the current undertakings aiming at the presentation of all of the Calderónian dramas as historical-critical editions[44] can take up the findings presented in this study.

# 5. Appendix

Abbreviations: T = Tragedy, C = Comedy, M = Mixed Cluster, U = Undefined Cluster

| Brief description and name of drama | Euklid Ward.D2 | Euklid Ward Sparse20 | POS Cosine | tf-idf Cosine |
|---|---|---|---|---|
| T1-A secreto agravio | M | T | C | C |
| T2-El alcalde de Zalamea | M | T | C | C |
| T3-El mágico prodigioso | T | T | T | T |
| T4-El mayor monstruo del mundo | T | M | T | T |
| T5-El médico de su honra | M | T | C | C |
| T6-El pintor de su deshonra | M | T | C | C |
| T7-El príncipe constante | T | T | T | T |
| T8-La devoción de la Cruz | T | T | T | T |
| T9-La hija del aire. Primera parte | T | T | T | T |
| T10-La hija del aire. Segunda parte | T | T | T | T |
| T11-La vida es sueño | T | T | T | T |
| T12-La gran Cenobia | T | T | T | U |
| T13-Las tres justicias en una | M | T | C | C |
| T14-Los cabellos de Absalon | T | T | T | T |
| T15-Saber del bien y del mal | T | T | T | U |
| C1-Casa con dos puertas mala es de guardar | M | C | C | C |
| C2-También hay duelo en las damas | C | C | C | C |
| C3-El encanto sin encanto | M | C | T | C |
| C4-Fuego de dios en el querer bien | C | C | C | C |
| C5-El astrólogo fingido | C | C | C | C |
| C6-El maestro de danzar | C | C | C | C |
| C7-La dama duende | M | C | C | C |

---

[44] A critical new edition of the complete body of *comedias* is in progress under the direction of Ignacio Arellano within the series *Biblioteca Aurea hispánica* from the Vervuert publishing house. Currently, however, only 21 titles have been published. This editing project can be seen as the most reliable textual basis; the editing principles are clarified in Arellano 2007. Additionally, the *Partes de las comedias*, which appeared during Calderón's lifetime, are available in a modern edition in six volumes through the Madrid-based publisher Fundación José Antonio de Castro, newly edited under the direction of Luis Iglesias Feijo.

| | | | | |
|---|---|---|---|---|
| C8-Los empeños de un acaso | C | C | C | C |
| C9-Mejor está que estaba | M | C | C | C |
| C10-Peor está que estaba | M | C | C | C |
| C11-Primero soy yo | C | C | C | C |
| C12-Mañanas de abril y mayo | C | C | C | C |
| C13-Antes que todo es mi dama | C | C | C | C |
| C14-No siempre lo peor es cierto | C | C | C | C |
| C15-Dicha y desdicha del nombre | C | C | C | C |
| Test1-Afectos de odio y amor | M | M | T | T |
| Test2-El galan fantasma | M | C | T | U |
| Test3-Las fortunas de Androméda y Perseo | M | M | U | T |
| Test4-Los dos amantes del cielo (T) | T | T | T | T |
| Test5-Amor, honor y poder (T) | T | T | T | T |
| Test6-La cisma de Ingalaterra (T) | T | T | T | T |
| Test7-En esta vida todo es verdad y todo mentira | M | M | U | T |
| Test8-La aurora en Copacabana | M | M | U | T |
| Test9-Las cadenas del demonio (T) | T | T | T | T |
| Test10-Amado y aborrecido | M | M | T | T |
| Test11-Amar después de la muerte o el Tuzaní de la Alpujarra | M | T | C | T |
| Test12-Las armas de la hermosura | M | M | T | T |
| Test13-Celos, aun del aire, matan | M | M | U | T |
| Test14-Darlo todo y no dar nada | M | M | T | T |
| Test15-Eco y Narciso | M | M | T | T |
| Test16-Fieras afemina amor | M | M | U | T |
| Test17-Luis Pérez el Gallego | M | T | T | C |
| Test18-El mayor encanto, amor (T) | T | T | T | T |
| Test19-La púrpura de la rosa | M | M | U | T |

| | | | | |
|---|---|---|---|---|
| Test20-El sitio de Breda | M | T | T | T |
| Test21-Nadie fíe su secreto | M | T | C | C |
| Test22-No hay burlas con el amor (C) | C | C | C | C |
| Test23-El escondido y la tapada (C) | C | C | C | C |
| Test24-No hay cosa como callar (C) | C | C | C | C |
| Test25-Las Manos Blancas No Ofenden | M | M | T | C |
| Test26-Con quien vengo, vengo (C) | C | C | C | C |
| Test27-Céfalo y Pocris (T) | T | T | T | T |
| Test28-La puente de Mantible (T) | T | T | T | T |
| Test29-El castillo de Lindabridis (T) | T | T | T | T |
| Test30-El monstruo de los jardines | M | M | U | T |
| Test31-La fiera el rayo y la piedra | M | M | U | T |
| Test32-Para vencer a amor, querer vencerle | M | T | T | T |
| Test33-Lances de amor y fortuna | T | T | T | U |
| Test34-Hombre pobre todo es trazas (C) | C | C | C | C |
| Test35-Judas Macabeo (T) | T | T | T | T |
| Test36-El alcaide de sí mismo (T) | T | T | T | C |
| Test37-El purgatorio de san Patricio (T) | T | T | T | T |
| Test38-La banda y la flor | M | T | T | U |
| Test39-Un castigo en tres venganzas | M | T | T | U |
| Test40-Bien vengas mal | C | C | T | C |
| Test41-Mañana será otro día (C) | C | C | C | C |
| Test42-La sibila del Oriente y gran reina de Sabá (T) | T | T | T | T |
| Test43-Argenis y Poliarco (T) | T | T | T | T |
| Test44-El jardin de Falerina | M | M | C | T |
| Test45-Los tres mayores prodigios (T) | T | T | T | T |
| Test46-Origen, pérdida y restauración de la Virgen del Sagrario (T) | T | T | T | T |

| | | | |
|---|---|---|---|
| Test47-La desdicha de la voz (C) | C | C | C | C |
| Test48-El secreto a voces | M | T | T | U |
| Test49-El Faetonte | M | M | C | T |
| Test50-La exaltación de la cruz (T) | T | T | T | T |
| Test51-El agua mansa (C) | C | C | C | C |
| Test52-La niña de Gómez Arias | M | T | C | T |
| Test53-Los hijos de la fortuna, Teágenes y Cariclea | M | M | T | T |
| Test54-Agradecer y no amar | M | M | T | C |
| Test55-Amigo amante y leal | M | C | C | C |
| Test56-El golfo de las sirenas | M | M | U | T |
| Test57-Gustos y disgustos son no más que imaginación | M | T | C | U |
| Test58-El acaso y el error | M | T | T | U |
| Test59-El José de las mujeres (T) | T | T | T | T |
| Test60-Los tres afectos de amor piedad desmayo y valor | M | M | T | T |
| Test61-Cada uno para sí | C | C | U | C |
| Test62-El conde Lucanor | M | M | T | U |
| Test63-Dar tiempo al tiempo (C) | C | C | C | C |
| Test64-Mujer, llora y vencerás | M | M | T | T |
| Test65-Cuál es mayor perfección, hermosura o discreción (C) | C | C | C | C |
| Test66-El laurel de Apolo | M | M | U | T |
| Test67-Ni amor se libra de amor | M | M | U | T |
| Test68-El mayor monstruo los celos | T | M | T | T |
| Test69-El postrer duelo de españa | M | C | C | C |
| Test70-El gran príncipe de Fez | M | M | T | T |
| Test71-Fineza contra fineza | M | M | U | T |
| Test72-El segundo Scipión | M | M | T | T |

| | | | | |
|---|---|---|---|---|
| Test73-La señora y la criada | M | T | T | U |
| Test74-Basta callar | M | C | T | U |
| Test75-De una causa dos efectos | M | T | T | T |
| Test76-Hado y divisa de Leonido y Marfisa | M | M | T | T |
| Test77-La estatua de Prometeo | M | M | U | T |
| Test78-Apolo y Climene | M | M | U | T |
| Test79-Duelos de amor y lealtad | M | M | T | T |
| Test80-Auristela y Lisidante | M | M | T | T |
| Test81-Cómo se comunican dos estrellas contrarias | M | T | T | C |
| Test82-La selva confusa | M | T | T | U |

# Bibliography

Ignacio Arellano: El Teatro de Corte y Calderón. In: Atti della Tavola Rotonda sulla Singolarità Storica e Estetica di »La púrpura de la rosa« di Calderón de la Barca. Ed. by María Luisa Tobar. Messina 2000, pp. 31–53. [Nachweis im GVK]

Ignacio Arellano: Editar a Calderón. Hacia una edición crítica de las comedias completas. Frankfurt / Main 2007. (= Comedias completas de Calderón, 5) [Nachweis im GVK]

Ignacio Arellano: Calderón y los géneros dramáticos, con otras cuestiones anejas. Honor, amor, legitimación política y autoridad de las taxonomías. In: Rilce. Revista de Filología Hispánica 34 (2018), pp. 100–126. DOI: 10.15581/008.34.1.100-26 [Nachweis im GVK]

Walter Benjamin: Ursprung des deutschen Trauerspiels. Frankfurt / Main 1978. (= Suhrkamp-Taschenbuch Wissenschaft, 225) [Nachweis im GVK]

Piotr Bojanowski / Edouard Grave / Armand Joulin / Tomas Mikolov: Enriching Word Vectors with Subword Information. In: Transactions of the Association for Computational Linguistics 5 (2017), pp. 135–146. PDF. [online]

John Andrew Bullinaria / Joseph P. Levy: Extracting Semantic Representations from Word Co-occurrence Statistics. A Computational Study. In: Behavior Research Methods 39 (2007), pp. 510–526. DOI: 10.3758/BF03193020 [Nachweis im GVK]

Pedro Calderón de la Barca: Obras completas. Textos íntegros según las primeras ediciones y los manuscritos autógrafos. Ed. by Ángel Valbuena Briones / Luis Astrana Marín. 3 vols. Madrid 1951–1956. [Nachweis im GVK]

Pedro Calderón de la Barca: Comedias y otras obras. Madrid 2007–2010. [Nachweis im GVK]

Miguel Campión Larumbe / Álvaro Cuéllar: 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. In: Talía. Revista de estudios teatrales 3 (2021), pp. 59–69. DOI: 10.5209/tret.74021

Enrica Cancelliere: Calderón e il Teatro di Corte. In: Atti della Tavola Rotonda sulla Singolarità Storica e Estetica di »La púrpura de la rosa« di Calderón de la Barca. Ed. by María Luisa Tobar. Messina 2000, pp. 55–76. [Nachweis im GVK]

María Esther Castro de Moux: Alquimia y gnosticismo en Fortunas de Andrómeda y Perseo de Calderón: In: Actas del V Congreso Internacional. Ed. by Christoph Strosetzki. (Asociación Internacional Siglo de Oro (AISO), Münster, 20.–24.07.1999) Frankfurt / Main 2001, pp. 319–330. [Nachweis im GVK]

Jorge Checa (Ed.): Pedro Calderón de la Barca: Amar después de la muerte. Edición y estudio. Kassel 2010. (= Teatro del Siglo de Oro / Ediciones críticas, 167) [Nachweis im GVK]

Erik Coenen: »La selva confusa« y »Cómo se comunican dos estrellas contrarias«: comedias gemelas. In: Revista de filología española 96 (2016), pp. 61–80. DOI: 10.3989/rfe.2016.03

Christophe Couderc: Le théâtre tragique au Siècle d'or. Cristóbal de Virués, Lope de Vega, Calderón de la Barca. Paris 2012. [Nachweis im GVK]

Álvaro Cuéllar: Stylometry and Spanish Golden Age Theatre: An Evaluation of Authorship Attribution in a Control Group of Undisputed Plays. In: Digital Stylistics in Romance Studies and Beyond. Ed. by Christof Schöch / José Calvo Tello / Ulrike Henny-Krahmer / Robert Hesselbach / Daniel Schlör. [Forthcoming]

Hanno Ehrlicher: Einführung in die spanische Literatur und Kultur des Siglo de Oro. Berlin 2012. [Nachweis im GVK]

Hanno Ehrlicher / Jörg Lehmann / Nils Reiter / Marcus Willand: La poética dramática desde una perspectiva cuantitativa: la obra de Calderón de la Barca. In: Revista de Humanidades Digitales 5 (2020), pp. 1–25. DOI: 10.5944/rhd.vol.5.2020.27716

Juan Manuel Escudero Baztán: Amor, honor y poder o el universo dramático de Calderón. Madrid et al. 2021. (= Comedias completas de Calderón, 24) [Nachweis im GVK]

Margaret Rich Greer: The Play of Power: Calderón's »Fieras afemina amor« and »La estatua de Prometeo«. In: Hispanic Review 56 (1988), issue 3, pp. 319–341. [Nachweis im GVK]

Matthew Jockers: Macroanalysis. Digital Methods & Literary History. Urbana, IL et al. 2013. [Nachweis im GVK]

Simon Kroll: Sonido y afecto en Calderón. Un estudio de las asonancias. Kassel 2022. [Nachweis im GVK]

Jörg Lehmann: Classification of Tragedies and Comedies in Calderón de la Barca's Comedias Nuevas [Data set]. In: zenodo.org. Version 1 from 20.06.2022. DOI: 10.5281/zenodo.6669603

Félix Lope de Vega: Arte nuevo de hacer comedias en este tiempo. Dirigido a la Academia de Madrid. Madrid 1621 [1609]. In: books.google.de. Original from la Biblioteca de Catalunya, digitized on 31.03.2010. [online]

Will Lowe: Towards a Theory of Semantic Space. Proceedings of the Annual Meeting of the Cognitive Science Society 23 (2001), pp. 576–581. [online]

Jesús G. Maestro: Los límites de una interpretación trágica y contemporánea del teatro calderonniano: El príncipe constante. In: Teatro calderoniano sobre el tablado: Calderón y su puesta en escena a través de los siglos. Ed. by Manfred Tietz. (Coloquio Anglogermano sobre Calderón, Firenze 10.–14.07.2002) Stuttgart 2003, pp. 285–327. (= Archivum Calderonianum, 10) [Nachweis im GVK]

Christopher D. Manning / Prabhakar Raghavan / Hinrich Schütze: Introduction to Information Retrieval. Cambridge, UK 2008. [Nachweis im GVK]

Tomas Mikolov / Ilya Sutskever / Kai Chen / Greg Corrado / Jeffrey Dean: Distributed Representations of Words and Phrases and Their Compositionality. In:Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems. Ed. by Chris Burges et al. (NeurIPS 26, Lake Tahoe, NV, 05.–10.12.2013), pp. 3111–3119. [online] [Nachweis im GVK]

Sean Papay / Sebastian Padó / Ngoc Thang Vu: Addressing Low-Resource Scenarios with Character-aware Embeddings. In: Subword and Character Level Models in NLP – proceedings of the second workshop. Ed. by Association for Computational Linguistics. (NAACL-HLT 16, New Orleans, 06.06.2018) Stroudsburg, PA, 2018, pp. 32–37. DOI: 10.18653/v1/W18-1204

Alexander Augustine Parker: Towards a Definition of Calderonian Tragedy. In: Bulletin of Hispanic Studies 39 (1962), pp. 222–237. [Nachweis im GVK]

Alexander Augustine Parker: The Mind and Art of Calderón. Essays on the Comedias. Ed. by Deborah Kong. Cambridge et al. 1988. [Nachweis im GVK]

Miriam A. Peña-Pimentel: El Gracioso en el Teatro de Calderón. Un Análisis desde las Humanidades Digitales. London / Ontario 2011. (= Electronic Thesis and Dissertation Repository, 307) [online]

Yves Peirsman / Dirk Geeraerts / Dirk Speelman: The Automatic Identification of Lexical Variation between Language Varieties. In: Natural Language Engineering 16 (2010), issue 4, pp. 469–491. DOI: 10.1017/S1351324910000161 [Nachweis im GVK]

Miriam A. Peña-Pimentel: Aplicación de mapas de tópicos al análisis semántico de algunas comedias de Calderón. In: Calderón virtual. Anuario calderoniano 5 (2012), pp. 115–130. [Nachweis im GVK]

Javier de la Rosa / Adriana Soto-Corominas / Juan Luis Suárez: The Role of Emotions in the Characters of Pedro Calderón de la Barca's autos sacramentales. In: Emotion and the Seduction of the Senses, Baroque to Neo-Baroque. Ed. by Lisa Beaven / Angela Ndalianis. (Conference, Melbourne, 27.–29.11.2013) Kalamazoo 2018, pp. 99–125. (= Studies in medieval and early modern culture, 59) [Nachweis im GVK]

Christof Schöch: Fine-Tuning our Stylometric Tools. Investigating Authorship and Genre in French Classical Drama. In: Digital Humanities Conference 2013. Hg. von European Association for Digital Humanities. (DH 2013, Lincoln, NE, 16.-19.07.2013) Lincoln, NE 2013. [Nachweis im GVK]

Christof Schöch: Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama. In: Digital Humanities Quarterly 11 (2017), pp. 1–53. [online]

Henry Wells Sullivan: Calderón in deutschen und niederen Landen. Eine dreihundertjährige Rezeptionsgeschichte. Berlin 2017. [Nachweis im GVK]

Henry Wells Sullivan: Tragic Drama in the Golden Age of Spain. Kassel 2018. (= Teatro del Siglo de Oro / Estudios de literatura, 133) [Nachweis im GVK]

María Luisa Tobar: Atti della Tavola Rotonda sulla Singolarità Storica e Estetica di »La púrpura de la rosa« di Calderón de la Barca. Messina 2000. [Nachweis im GVK]

Peter D. Turney / Patrick Pantel: From Frequency to Meaning: Vector Space Models of Semantics. In: Journal of Artificial Intelligence Research 37 (2010), pp. 141–188. DOI: 10.1613/jair.2934 [Nachweis im GVK]

Ángel Valbuena Prat: Historia de la literatura española. 4 vols. 3rd edition. Barcelona 1950. Vol. 2: Los Siglos de oro, pp. 479–571. [Nachweis im GVK]

Joe H. Ward: Hierarchical Grouping to Optimize an Objective Function. In: Journal of the American Statistical Association 58 (1963), pp. 236–244. [Nachweis im GVK]

Marcus Willand / Nils Reiter: Geschlecht und Gattung. Digitale Analysen von Kleists ›Familie Schroffenstein‹. In: Kleist-Jahrbuch 2017. Ed. by Andrea Allerkamp / Günter Blamberger / Ingo Breuer / Barbara Gribnitz / Hannah Lotte Lund / Martin Roussel. Stuttgart 2017, pp. 177–195. [Nachweis im GVK]

# List of Figures and Tables

Fig. 1: Ward.D2 clustering of 112 Calderónian Comedias. [Lehmann 2022]

Fig. 2: Ward.D2 clustering of 112 Calderónian Comedias. Euclidian distance on the basis of a sparsity of 20%. [Lehmann 2022]