Contribution for the Zeitschrift für digitale Geisteswissenschaften: Calderón-Comedia Nueva

Corresponding author: Dr. Jörg Lehmann
E-Mail: joerg.lehmann@romanistik.uni-tuebingen.de
Romanisches Seminar
Eberhard-Karls-Universität Tübingen
Wilhelmstraße 50
72074 Tübingen
ORCID-ID: 0000-0003-1334-9693
DNB: 1054732310
Twitter: @jrglmn

Second author: Prof. Dr. Sebastian Padó
E-Mail: sebastian.pado@ims.uni-stuttgart.de
Institut für maschinelle Sprachverarbeitung
Abteilung Theoretische Computerlinguistik
Universität Stuttgart
Pfaffenwaldring 5b
70569 Stuttgart
ORCID-ID: 0000-0002-7529-6825
DNB: 1033924393
Twitter: @nlpado

Title: Classification of Tragedies and Comedies in Calderón de la Barca

Abstract

In this study, the classification of 64 dramas written by Calderón de la Barca is accomplished by applying procedures established by distributional semantics. The objective is to distinguish between comedies and tragedies within this drama corpus. Fifteen of these *comedias nuevas* have already been classified by qualitative researchers as tragedies and comedies, respectively; for another 34 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. Based on the words identified via the log-likelihood probability distribution, the 64 dramas can be divided into two subgroups using the K-nearest neighbor method.The outcome shows that the explored methods identify tragedies with greater accuracy than comedies. With regard to distributional semantics, however, 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 mitologicas*.

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 works include 84 Corpus Christi plays (*autos sacramentales*), 110 *comedias* and 41 short pieces (*teatro cómico breve*). 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, upon whose pragmatic rules Calderón generally oriented himself, despite any 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, and the tragedy as pertaining to wealthy landowners or members of the royal family and being based on historic events. Furthermore, Lope characterizes the *comedia nueva* as a mixture of comedic and tragic elements, thus referring to the combination of both dramatic genres.[4] Therewith, 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 "fortuitous" 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 intensively on the tragedies of the Spanish Golden Age and practically implemented his theoretical aspirations 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 »Spanish Theater« (Vol. I: 1803, Vol. II: 1809) and examined Calderón in great detail in his

[1] Calderón de la Barca 1951–1956. This publication, however, does not conform to the standards of a historio-critical edition.
[2] Sullivan 2018, p. 33.
[3] Lope de Vega 1621.
[4] This may be considered a reference to a third genre, which has received little attention up to now in the research. Compared to Couderc 2012, p. 65–75 and 102–109.
[5] Comprehensively in detail Sullivan 2017.

»Vorlesungen über dramatische Kunst und Literatur (Readings 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 goes into detail about 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-twentieth 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 the antique, which, since the time of Aristotle have been based on the clear separation of comedy and tragedy; however they 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 T. Sullivan among others) was intensely occupied with the Calderón 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, from a qualitative perspective, to identify 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), characteristics of the plot (unfair judgements or death of the protagonist), attributes of reception (creation of *eleos* and *pathos* or cathartic endings), or the formulation of exclusionary criteria (for instance, themes such as redemption and damnation or martyr dramas).[9] Within the framework of these criteria, Sullivan was able to identify at least 14 tragedies in the complete works of the Calderónian *comedias.*

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 publishers from the Aguilar edition – was never carried out comprehensively:[10] What researcher is prepared to study and classify 110 dramas? At the same time, it is evident that just this sort of written work is suitable for

[6] Benjamin 1978.

[7] Compare here the introduction by Ángel Valbuena Briones, in: Calderón de la Barca 1951, p. 9–34.

[8] Comp. Maestro 2003 and also the discussion by Arellano 2018 on the limits of compiling taxonomies.

[9] Sullivan 2018, p. 362–364.

[10] An attempt at this is being made by the portal #Calderón Digital [http://calderondigital.tespasiglodeoro.it/], by which around 80 of Calderón's written texts can be filtered according to genre characteristics; the researchers responsible for these classifications are also included.

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* is rendered impossible, as they still remain partially unavailable in an electronic form. 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 they quite obviously lend themselves to the examination of structural similarities among works in a particular genre or differences between dramas of varying genres.[11] 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[12] represents an attempt, based on at least 64 *comedias,* not only to critically assess the validity of the distinction between the comedy and the tragedy, but also to exercise the methodical possibilities made available by the *digital humanities'* application of *distributional semantics* procedures for this problem.[13] Because, thus far, only a small portion of the Calderónian *comedias* have been studied, and the majority of them remain entirely unexplored, it is to be expected that the proven methods can deliver important indications for the classification of each play which has yet to be thoroughly analyzed.

2. Methodical Procedures

Methodical Basis.

Nowadays, the concept of distributional semantics is widespread in the realm of computer linguistics. Basically, it is assumed that the meaning of a word is established according to how much it is used and how often it is coupled with other words within a specific context. Presumably, words and documents are represented in a highly dimensional vector space and bound by semantic relationships through similarities within that space. For the representation of documents, the frequency (absolute or relative) of the words in each document are set into matrices, whereby each word creates a column of the matrix, and every document a line. The frequency data is found in the corresponding segments of the matrix; 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.[14] In order to represent the meanings of words, the same kind of matrix is created, with the targeted terms forming lines and contextual words forming segments. On the basis of such matrices, the distances between single words or texts are computed, compared to each other, cumulated into groups through clustering methods and visualized. As a rule, these very large matrices are sparse so that they can be reduced to a much smaller number of dimensions in order to serve as a basis for distance or similarity matrices. The resulting low dimensional vectors are often

[11] For example Peña-Pimentel 2011, 2012. de la Rosa et al. 2018. Ehrlicher et al. 2020.

[12] This study arose as a part of the project "QUOTE". "Comprehensive Modelling of Speaking in Prose Texts", sponsored by the German Research Community (Deutsche Forschungsgemeinschaft, project nr. 350397899). The authors thank Prof. Dr. Hanno Ehrlicher (University of Tübingen), who commented on the first version of the article.

[13] Comparable studies on classical French drama have been thus far presented by, for instance, Christof Schöch 2017 and 2013, who approached the subject with *topic modeling a*nd stylometric methods. For stylometric analysis of dramas in the *siglo de oro* comp. In particular, Campión Larumbel and Cuéllar González 2021, and Cuéllar González 2022.

[14] Comp. Lowe 2001 for details.

referred to as *word* or *document embeddings* and are 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 purely a technical requirement and alters little on the underlying intention.[15]

The reason for choosing the distributional approach for the work at hand provides for the assumed outcome, that comedies and tragedies—in accordance with the treatment of each of the different themes—can be differentiated through varied choices of words and through each of the various groupings of these words. Simply put, it can be expected that in Calderónian tragedies, terms such as honor, power and death strongly correlate, while the comedies tend to combine words like love, disguise and jealousy. This is quite obviously an approach that represents a rough simplification—narrative patterns or plot structures, however, cannot be characterized in this manner. At the same time, the wide success of word-based approaches such as topic-models and common methods for author recognition demonstrates that simple co-occurence analyses allow for surprisingly deep understandings even in literary texts.

Data Basis.

Beginning with the 14 tragedies identified by Sullivan, yet another was added to the examined texts, which had apparently remained unknown to him.[16] Fifteen further dramas, which were identified as comedies during the research,[17] make up the counterpart to the tragedies in this body of work. The other 34 Calderónian *comedias* represent those which are currently available as full digital texts in modernized and normalized Spanish.[18] Although digital copies of all the 17th century publications of the Calderónian *comedias* exist, the OCR and transcriptions of these plays into modern Spanish were not carried out for pragmatic reasons. The spoken texts of the *dramatis personae* were extracted from all 64 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 34 plays were marked »Test« and also numbered.[19]

Research Goal.

In the absence of suitably large bodies of dramatic works, the classification of genre with *word* or *document embeddings* is still relatively new.[20] Thus, the goal of our study is to explore various methods and combinations thereof, and to compare the results. We will compare, henceforth, four valuations, which all follow the same general unobserved schemes: (a) pre-filtering of the vocabulary; (b) calculation of *document embeddings*, and,

[15] Jockers gives a short introduction 2013, p. 63–67.

[16] Comp.recently to this identification. Escudero Baztán 2021, p. 21.

[17] Comp. additionally Calderón de la Barca 1951; Escudero Baztán 2021; Ehrlicher 2012; Maestro 2003; Parker 1988; Peña-Pimentel 2011; Tobar 2000; Valbuena Prat 1950.

[18] For the most part, these dramas are available under the portal: #Cervantes Virtual [http://www.cervantesvirtual.com/] and the #Association for Hispanic Classical Theater [http://www.comedias.org/]. A current overview of all sources can be found at: #Estilometría aplicada al Teatro del Siglo de Oro [http://etso.es/]. Because diacritical symbols used in modern Spanish can be used according to context, the spelling of certain terms may vary. (ex.: solos / solós).

[19] Comp. Under attachment in which this seal was removed and the results of the applied methods are presented.

[20] One exception compiled the study by 2017, comp. here p. 190–194.

if applicable, dimension reduction; (c) clustering of *embeddings*; (d) visualization und evaluation. The aforementioned corpus provides us with an excellent basis, as the categories are known in about half of the plays, but not in the other half: by this means, the quality of the process can simultaneously be reviewed (on the basis of the known categories) and findings on the new dramas can be obtained. We find this type of methodical comparison important, because it is known that the findings from unobserved distributional methods depend heavily on the parametrization of the process.[21]

Practical Application.
All evaluations were made with the statistic 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 body of text to be processed in other packages. As was shown in the course of exploration, only a small number of Spanish stop words were retained in the quanteda package (just 308). 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.[22] Furthermore, the analysis of the tf-idf algorithm in particular, showed that the grouping results were quite negatively affected by names of characters and places within the text, as these elements of speech, due to similarities between documents, tend to reflect idiosyncrasies of single pieces rather than stereotypical genre characteristics. These proper names were likewise—first and foremost through the list of *dramatis personae*—compiled and removed from the texts. As a rule, the relative 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.

3. Results

Experiment 0.
In one of the first explorations, we combined the body of text with the Skip Gram standard method for *word embeddings* in a matrix[23] in order to determine whether *word embeddings* could tell us anything discernible about the text and which word pairings within the entire body of 64 dramas exhibited the highest amount of similarities. The resulting matrix was reduced to 1,000 terms with the highest log-likelihood probability distribution and the cosine similarity between all vector pairs was calculated. The cosine similarity measures the cosine of the angle between two vectors and determines whether they more or less point in the same direction within the high dimensional space. In this way, it can be determined whether two terms are found in similar contexts, whereby the values range between 0 and 1 and where cosine similarity values more inclined towards 1 exhibit similar

[21] Turney and Pantel 2010; Bullinaria and Levy 2007.
[22] These word lists are documented in the R code, which was published together with the body of dramas in the DARIAH-DE repository under [https://doi.org/10.20375/0000-000E-6729-1] Comp. Lehmann 2021.
[23] Mikolov et al. 2013.

contexts.

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 0.97, 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 impressive, in that, by using the Skip Gram algorithm, central themes in the Calderónian *comedias* could be identified, even though they deal with the interfaces of social conventions (honor) and individuality (taste, soul, fear, social or actual death).

Conversely, word pairings like »honor« and »poder« (honor and power) (0.58), »amores« and »agravios« (love and infidelity, each in plural form) (0.65) »gracia« and »corte« (grace and court) (0.63) or »gracia« and »culpa« (grace and guilt) (0.51) 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 and reputation), »muere« and »sepulcro« (he/ she/ it dies and grave), »muerte« and »engaño« (death and deceit), »mueran« and »suerte« (they may die and fate), »amores« and honra« (love, reputation) and also »mentira« and »gracia« (lie and grace). First and foremost, it is apparent that the central themes in Calderón's works (»amor«, honor y poder«[24] – 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.

Experiment 1.

With the first experiment, our goal was to be able to weigh the unobserved *document embedding* processes according to their validity. 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 distance algorithm[25] 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 algorithm, 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.

[24] Comp. Escudero Baztán 2021.

[25] Ward 1963.

The first method represented a conservative approach: only the 1,056 words with a frequency > 70 and appearing in at least half of the documents were included. The document word matrix was filled with mere frequencies; no dimension reduction took place. The grouping was carried out through a clustering with the Ward.D2 distance algorithm. Image 1 shows the resulting dendrogram.

Figure 1: Ward.D2 Clustering of 64 Calderón dramas. [Lehmann 2021]

Read from left to right, the first cluster represents a pure comedy cluster which includes only twelve dramas; yet eight of these had already been distinguished as comedies. The cluster all the way to the right depicts a pure tragedy cluster; here 18 dramas are included, of which nine had already been classified as tragedies. In the middle, between both clusters, two additional clusters are shown which must be described as mixed clusters, as each contain both tragedies and comedies. Together, the middle clusters contain more than half of the plays, namely 34 works. This process seems, with regard to the main research question, not to be especially effective, as only 13 of the 30 previously marked dramas (or 43%) were allocated in a clear fashion, while many comedies and tragedies mutually appeared in the clusters. However, the still relatively high dimensionality of the *document embeddings* hinders a failure analysis.
The goal of the second process is to create a low dimensional representation that is easier to interpret, in order to gain more insight. First, only terms which appear in at least 80% of all of the documents (50 Plays) are retained. This reduces the number of terms to a more compact total of 471. Again, a frequency based word-document matrix is established and normalized with multi-dimensional scaling, whereby each of the remaining terms in each drama are 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.

Figure 2: Ward.D2 Clustering of 64 Calderón dramas, Euclidian distance based on a *sparsity* of 20%. [Lehmann 2021]

The dendrogram illustrates three clusters: In the first cluster to the left, twelve comedies and five further dramas appear. The cluster on the right contains 14 tragedies and, likewise, 14 dramas of unknown classification. The cluster in the middle is mixed; it contains three comedies (C3: »El encanto sin encanto«, C4: »El Faetonte«, C5: »El jardín de Falerina«), one tragedy (T4: »El mayor monstruo del mundo«) and 15 additional dramas of unknown classification. Through this process, which only deals with 471 words, 26 of 30 classified dramas, or 87%, were correctly allocated.[26]
A further result of both procedures, in which the fundamental matrices are reduced on the basis of word frequencies, is that a transitional segment is established between tragedy and comedy. This observation presents us with the question of whether it would be more

[26] Basically, we attempted to alter only one parameter between each of the analyses, thus using the Euclidian distance. As an alternative, during the second experiment, 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 categorized.

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 regards 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 honorific 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.[27]

One possible basic 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 content for classification.[28] 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.

Figure 3: Ward.D2 Clustering of 64 Calderónian plays, cosine similarities based on verbs, nouns and adjectives. [Lehmann 2021]

The first cluster to the left (comedies) contains twelve comedies, two tragedies (T2: »El alcalde de Zalamea«; T6: »El pintor de su deshonra«) and seven additional plays. The cluster to the right is purely a tragedy cluster, however it contains ten tragedies and nine additional plays. In the middle between these two categories is a mixed cluster, containing three comedies, three tragedies and eighteen additional plays. With regard to the plays identified thus far as tragedies and comedies, 73% of these dramas were correctly categorized.[29]

Taking into consideration the previously tested methods, it seems advisable to focus on every term that carries meaning, thus leading to a differentiation between the categories. The fourth method we tried was based on the tf-idf algorithm, the underlying method for which is *text mining* for degrees of common association, whereby terms can be evaluated for their significance within a document or body of work. With the tf-idf algorithm the importance of one of each term per document is calculated; the frequency of appearance of each term (*term frequency, tf*) is thereby multiplied by the inverse document frequency (*inverse document frequency, idf*). The latter depends, not on individual documents, but

[27] This possibility was already mentioned by Greer 1988 in an example from “Fieras afemina Amor”.

[28] This kind of method was used by Willand and Reiter 2017, p. 191f.

[29] As an alternative, an normalized matrix was established and a Ward.D2-Clustering based on the Euclidian distance was carried out. The results are slightly better, each comedy and tragedy was assigned to the appropriate cluster, so that all in all, 80% precision in classification was reached. However, five clusters resulted altogether, being that three individual „test“ plays were each identified in their own clusters. Compare that to the R-Code in the data publication.

rather, on the total number of all documents in the corpus. In this way, the tf-idf algorithm considers the relative significance of words which appear frequently in the text to determine how relevant the term is for a document within the body of work. 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.

Figure 4: Ward.D2 clustering of 64 Calderón plays. Cosine similarity based on tf-idf data. [Lehmann 2021]

This image shows three clusters: the first one to the left can best be described as a comedy cluster. In addition to thirteen comedies, however, it also contains four tragedies (T1: »A secreto agravio, secreta venganza«; T5: »El médico de su honra«;[30] T6: »El pintor de su deshonra«; T13: »La gran Cenobia«) and nine other dramas. The one on the right, with twelve tragedies, the comedy C5 »El jardin de Falerina« and fifteen further plays, can be considered a tragedy cluster. The smallest one in the middle is not well defined; it contains only one drama clearly identified as a comedy and eleven others which remain unclassified. In comparison with the dramas already identified as tragedies or comedies, this result shows that 26 of 30 tragedies and, respectively, comedies have been classified correctly; this correlates to a recognition rate of 77%.[31] Given the background of hitherto targeted outcomes, this would seem to be satisfactory, however, this process did not completely classify the comedies and tragedies with the same qualitative standards used by researchers. Furthermore, as was the case in the previously tried methods, there is a third cluster representing a transitional area between the two categories containing only one drama classified as a comedy.

The four methods explored here differentiate, for one thing, through the establishment of the main corpus of work and for another, through the choice of distance or similarity matrices. Three of the four generated robust to good results, but the process employing the strongest matrix reduction produced the best findings. In none of the cases, however, was the classification through clustering consistent with that of researchers applying qualitative analyses.

Experiment 2.

In a second experiment, contrastive vocabulary analyses were carried out. A vocabulary analysis is superimposed on the four previous methods and evaluates the word lists upon which the foundations of clusters are based; additionally, the log-likelihood probability distribution for two subgroups is calculated, which contain only dramas classified as comedies or tragedies. In this way, the 200 words with the highest log-likelihood values for each subgroup can be determined and the results can be compared (contrastive vocabulary analysis with *word embeddings*).

[30] This outcome is especially interesting, because, according to Couderc 2012, p. 104 both dramas can be described as tragicomedies and “A secreto agravio, secreta venganza” is the only play by Calderón which uses the term “tragicomedy” in the spoken text.

[31] Here, alternatively, a Ward.D2 clustering was also carried out based on the Euclidian distance. The resulting classification accuracy of 77% is identical to the subsequent values of the cosine similarity (23 of 30 plays were correctly assessed).

In the first procedure, Ward.D2 clustering based on the Euclidian distance between normalized word frequencies was carried out; only the first and the fourth 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 fifteen terms with the highest probability margin for each were selected. These fifteen 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«, »papel«, »calle«, »dama«, »puerta«, »cuarto«, padre«, »caballero«, »señora, »saber«, »honor«, »criado«, »amigo« and »hermano« (esquire, house, paper, street, lady, door, room, father, gentleman, woman, knowledge, honor, servant, friend and brother) appeared. Except for the word „honor", this word list does not seem to be significantly distinctive of comedies. For the tragedy cluster, however, the words »rey«, »muerte«, »dios, »hoy, »cielo«, »señor«, »sol«, »valor«, »rigor«, »alma«, »sangre«, »reina«, »gran«, »quiero« and vida« (king, death, gods, today, heaven, mister, sun, value / valor, severity, soul, blood, queen, grand, I want and life) were especially frequent. At any rate, people of high social standing, death, soul, valor and blood stand out as being characteristic terms relating to these storylines.

The 471 words selected for their sparsity of 20% enable a preview of terms which carry a strong distinction in classifying comedies and tragedies. For the comedy cluster, meaningful terms like »don«, »casa«, »dama«, »puerta«, »cuarto«, »calle«, »señora«, »papel«, »padre«, »honor«, »bien«, »noche«, »cuidado«, »caballero« and »hombre« (gift, house, lady, door, room, street, woman, paper, father, honor, good, night, care, knight and man) are present. For the tragedy cluster, words like »rey«, »señor«, »dios«, »muerte«, »hoy«, »cielo«, »alma«, »sangre«, »rigor«, »mundo«, »ocasión«, »viento«, »sol«, »quiero« and »vida« (king, mister, God, death, today, heaven, soul, blood, severity, world, occasion, wind, sun, I want and life) appear. Primarily, 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 precision for classification in 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 "to be" and "to have" »ser« and haber«, with a large gap followed by a row of 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« oder »don« (mister, 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 good.

In the fourth procedure—based on the tf-idf matrix—an approach analogue to methods 1 and 2 is applied. The fifteen terms that carry the highest probability margin within the comedy cluster are: »don«, »doña«, »tapada«, »duque«, »sólo«, »criada«, »papel«, »cuarto«, »casa«, »criado«, »anoche«, »parque«, »aposento«, »hermana« and »máscara« (esquire, lady, veil, duke, only, maid, paper, room, house, last night, park, chamber, sister and mask). In the tragedy cluster, terms such as »conde«, »fez«, »rey«,

»cristianos«, »villa«, »castillo«, »ejército«, »senado«, »emperador«, »galera«, »fué«, »soldados«, »puente«, »capitán« and »cruz« (count, fez, king, Christians, manor, castle, army, senate, king, galley, was / were, soldiers, bridge, captain and cross) 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, the aristocratic descent of the protagonists as well as military and Christian themes.

Apart from the word lists that provide the foundation for the four different clustering methods, and terms through which their degrees of probability are determined, it is also of interest to see how robustly the results may be estimated if they are each applied to the basis of clusters which do not include comedies or tragedies exclusively. Thus, in the next step, this discriminative word list, generated by means of a larger body of work, and the context in which the terms appear, will be put to the test. With this objective, the body of plays identified as comedies or tragedies will be expanded and two somewhat larger subgroups generated, in which the results of the previous four procedures can be compared. From the dramas hitherto marked as »Test«, six will be chosen which were unanimously or predominantly identified as being either »tragedy« or »comedy«. This was easily carried out for the tragedies, since five of the dramas were consistently classified in all four of the procedures as such. One additional drama was classified as a tragedy in three of the four previous procedures, and the editor of the most recent historical-critical edition has also classified it in this way.[32] In regard to the comedies, however, the identification of additional plays is not as easy. Not one of these plays was consistently classified as a comedy through the methods tried thus far. Regarding other dramas, which were identified as comedies in three procedures, no secondary literature exists which would support this evaluation. For this reason, six plays are tentatively chosen, which were included in the collection of comedies by the editors of the Aguilar edition and, where possible, their classification can also be corroborated by secondary literature.[33] In this manner, two new subgroups are now generated, one for tragedies and one for comedies, each containing 21 plays.[34] 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, desdichas, desengaño, favor, feliz, joya, juego, loco, máscara, secreto, suceso, tapada, tristeza, vestido – mistress, girlfriend, misfortune, disappointment, favor, joyful,

[32] Comp. Checa 2010, p. 13.

[33] Assess as comedy: for “El escondido y la tapada” comp. Escudero Baztán 2021, who described this drama as “comedia de capa y espada” (p. 63). For “No hay cosa como callar” comp. Parker 1988, who understood this play as being a “comedy of intrigue” (p. 181). Regarding “Las manos blancas no ofenden” comp. Valbuena Prat 1950, who counts this play amongst “obras exclusivamente cómicas” (p. 541), an estimation which was not corroborated by this procedure.

[34] Comp. for a comparative method Peirsman et al. 2010.

jewelry, game, crazy, mask, secret, event, deception, sadness or disguise), typical indications relating to the mythological background of the comedies (cristales, deidad, demonio, estatua, fiera, ninfas – crystals, deity, demon, statue, monster or nymph) and also the appearance of some rather surprising terms, (like disgusto, enemigo, pendencia, razón or saber – disgust, enemy, brawl, reason or knowledge).

By contrast, among the tragedies we find references to the (mostly high) standing of the characters (consejo, convento, corona, emperador, esclavo, infanta, infante, majestad, reina, reinar, reino, rey, tirano, villanos – counsel, cloister, crown, emperor, slave, infant, infanta, highness, authority, queen, ruling, kingdom, king, tyrant or villain), the contents of the plot (agravio, cristo, desdichado, esperanza, gloria, hermosura, laurel, lealtad, ley, libertad, morir, poder, salud, sangre, traíción, traidor, triste, triunfo, venganza – defamation, Christ, misery, hope, fame, beauty, laurel, devotion, rights / law, freedom, dying, power, health, blood, treason, traitor, sad, triumph or revenge) and a few surprises as well (ciencia / ciencias, enamorado, sueño – science / s, enamored or dream). Altogether, the word lists determined by the two subgroups and the log-likelihood outline the contents of the comedies and tragedies much more precisely than the word lists based on each cluster and upon which the rate of probability is established.

In order to make a text-based comparison of the 70 words appearing in both genres, we calculate *word embeddings* for these terms separately for each subgroup. Our goal is to characterize how these terms are used differently in each genre. For this purpose, we used the embedding method fastText[35] and the R-Package of the same name. In each subgroup, the 10 k-nearest neighbor terms of interest are established, in order that each word which was identified as pertaining to both genres is visible, along with the terms found closest to it within the text. The fastText method, developed by Facebook's AI Research Laboratory, calculates, much like the better-known Skip Gram method, *word embeddings,* with the objective of allocating words which are close together and often co-occurring with the most approximate *embeddings* possible. 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.[36]

As a contrasting distinction of the terms in each subgroup, we will illustrate in the following the ten k-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 »amistad« (friendship). In other words, both terms are used in comedies and tragedies, but within completely different contexts according to each. Again, it becomes apparent that the terms „honor“ and „friendship“ 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

[35] Bojanowski et al. 2017.

[36] Papay et al. 2018.

associated with possible death.

Comedia	Tragedia
honor	
hablado 0.92 (spoken)	satisfacer 0.79 (satisfying)
dado 0.88 (given)	honrar 0.77 (honoring)
prado 0.88 (meadow)	tratar 0.73 (treating)
enseñado 0.88 (learned)	remediar 0.71 (remedying)
hallado 0.88 (found)	satisfecho 0.71 (satisfied)
cercado 0.86 (enclosed)	favor 0.71 (liking)
honrado 0.86 (honored)	matar 0.71 (killing)
descalabrado 0.86 (hurt)	jugar 0.70 (playing)
causado 0.86 (caused)	castigar 0.69 (punishing)
pecado 0.86 (sinned)	faltar 0.68 (missing)
amistad (friendship)	
avisad 0.84 (warning)	amistades 0.77 (friendships)
podéis 0.77 (can)	dificultad 0.75 (difficulties)
dad 0.77 (giving)	amigo 0.74 (friend)
necedad 0.77 (stupidity)	determino 0.73 (determining)
podréis 0.77 (may)	justamente 0.72 (just)
quedéis 0.76 (stay)	libertad 0.72 (liberty)
soltad 0.76 (releasing)	mitad 0.71 (half)
novedad 0.76 (news)	testigo 0.71 (witness)
oiréis 0.75 (examining)	ingratitud 0.71 (ingratitude)
prosigáis 0.75 (continuing)	satisfación 0.71 (satisfaction)

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. 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 »justicia«, »fineza« and »muera« (justice, kindness, he/ she/ it dies) 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 mortal danger and the realm of divine providence and justice.

Comedia	Tragedia
justicia (justice)	
noticia 0.77 (notice)	**noticia** 0.81 (news)
justa 0.76 (just)	**justa** 0.81 (just)
malicia 0.71 (malice)	audiencia 0.80 (audience)
hidalguía 0.71 (generosity)	jurisdicción 0.80 (jurisdiction)
codicia 0.71 (greed)	traición 0.80 (treason)
gallarda 0.71 (pride)	paciencia 0.79 (patience)
galería 0.70 (gallery)	instancia 0.79 (instance)
idioma 0.70 (language)	prevención 0.79 (prevention)

góndola 0.70 (gondola) persuadida 0.70(persuaded)	providencia 0.78 (prediction) prudencia 0.78 (prudence)
fineza (nicety)	
infausta 0.80 (hapless) fama 0.80 (fame) firmeza 0.80 (firmness) grandeza 0.80 (greatness) finezas 0.79 (charities) amanezca 0.79 (fading) confianza 0.78 (confiding) necia 0.77 (foolishness) **nobleza** 0.77 (aristocracy) fingida 0.76 (pretending)	fiereza 0.77 (savagery) pieza 0.75 (piece) fortaleza 0.75 (strength) firmeza 0.75 (firmness) empieza 0.73 (begins) **nobleza** 0.73 (aristocracy) sutileza 0.72 (subtlety) alteza 0.72 (gratefulness) mudanza 0.72 (move) zalamea 0.72 (Zalamea)
muera (he / she / it dies)	
muriera 0.87 (dying) muestra 0.85 (showing) viviera 0.84 (living) viera 0.84 (seeing) defuera 0.84 (dying) verdadera 0.83 (genuine) **manera** 0.83 (way) entera 0.83 (entire) fiera 0.83 (monster) llora 0.82 (crying)	**muriera** 0.91 (dying) muerta 0.89 (dead) mísera 0.87 (miserable) matara 0.86 (killing) **manera** 0.86 (way) mueva 0.85 (moving) materia 0.85 (matter) magia 0.83 (magic) afuera 0.83 (outside) muda 0.83 (mute)

However, other terms clearly show overlaps with regards to the k-nearest neighbor terms; for instance, »celos«, »gusto« or »hado« (jealousy, taste or fate) each share four or five k-nearest neighbor terms within the ten words in the selection.

Comedia	Tragedia
celos (jealousy)	
celosos 0.92 (jealous) **cielos** 0.89 (heaven) **recelos** 0.89 (doubt) duelos 0.84 (duel) solos 0.83 (alone) **dellos** 0.83 (from them) filos 0.82 (cutting) **desvelos** 0.81 (care) testigos 0.81 (witness) desconsuelos 0.81 (hopelessness)	**dellos** 0.91 (from them) **recelos** 0.90 (doubt) cuellos 0.89 (necks) **cielos** 0.87 (heaven) cabellos 0.86 (hair) caballos 0.85 (horse) **desvelos** 0.85 (care) regalos 0.83 (gifts) ojos 0.83 (eyes) amenos 0.82 (pleasant)
gusto (taste)	
gano 0.84 (gaining) **justo** 0.83 (just) **disgusto** 0.82 (disgust) injusto 0.81 (unjust) susto 0.81 (frightening) misterio 0.81 (secret) **visto** 0.81 (seen)	**justo** 0.91 (just) augusto 0.88 (grateful) preciso 0.87 (precise) **visto** 0.84 (seen) **disgusto** 0.83 (disgust) presto 0.83 (fast) atrevo 0.82 (trusting)

cristo 0.81 (Christ)	precio 0.82 (prize)
llano 0.79 (plain)	agosto 0.81 (August)
gustosa 0.78 (enjoyable)	susto 0.81 (scare)
hado (fate)	
hablado 0.92 (spoken)	**dado** 0.88 (given)
dado 0.88 (given)	**hablado** 0.85 (spoken)
prado 0.88 (meadow)	echado 0.85 (thrown)
enseñado 0.88 (learned)	adorado 0.85 (adored)
hallado 0.88 (found)	enfado 0.85 (worshipped)
cercado 0.86 (enclosed)	atado 0.85 (bound)
honrado 0.86 (honorable)	prado 0.84 (meadow)
descalabrado 0.86 (hurt)	apartado 0.84 (section)
causado 0.86 (caused)	sañudo 0.83 (bitter)
pecado 0.86 (sinned)	**hallado** 0.83 (found)

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.

Experiment 3.

Consequently, based on the 200 most informative words for each subgroup, a (partially) observed method of classification is carried out in order to separate the entire body of work into a binary selection of comedies and tragedies. For the classification of the dramas through their k-nearest neighbor terms, a classifier is not trained as it is with other procedures, but rather, the k-nearest neighbor outcomes for each play are calculated with unknown classifications. The predictions are observed about which classification is more dominant in the neighboring term list for each test document and each is separated accordingly into its most dominant classification.

First, before we carry out this procedure, the usual steps must be taken for preprocessing—a document word matrix is generated containing a total of 330 terms each, exclusively determined through the log-likelihood probability of each subgroup (130 discriminative words for each subgroup + 70 terms which are identical in both subgroups). Because 30 of the documents have known classifications as “comedy” or “tragedy”, they will be used as a training set. For the remaining 34 *comedias* (or the test set), the classification for each will be calculated, in that the five k-nearest neighbor documents serve to discern which group they belong to. The results are documented in the attachment. Finally, the matching precision for classification is evaluated, using the classification of those twelve dramas from the test set through which both subgroups, made up of 21 comedies and tragedies each, were generated. This matching precision ran up to 83%, as ten of the twelve *comedias* were attributed to the same classification through the KNN method as they had been by the top qualitative researchers. This precision is slightly worse than the best unobserved method from experiment 1 (whereby, it should be noted that the interpretation of the clusters from experiment 1 also required forehand knowledge of the categorization of plays); this underscores, once again, the importance of an appropriate choice of parameters when implementing unobserved methods. From a qualitative viewpoint, one may ask whether it is at all sensible to

completely divide the examined body of work into two parts; the classification results could just as well be consigned to the intermediate area between comedies and tragedies and identified under the less observed classification of the tragicomedy.
A further result of the classification through the k-nearest neighbor documents was that all six tragedies presumed as such during the previously used methods, were once more identified thusly, however this was not the case with the six comedies. This result again points to the conclusion that apparently tragedies are more clearly classified through methods of distributional semantics than are comedies. The result of the k-nearest neighbor method is subsequently utilized for a final visualization: the dramas are each represented by the 330 terms with the highest log-likelihood values. The Euclidian distance matrix of these representations are then reduced into two dimensions with a scaling process. In the resulting figure 5, both classifications can be manually approximated as ellipses.

Figure 5: Euclidian distances between 64 dramas based on 330 log-likelihood words. [Lehmann 2021]

First, we see two distinct groups of comedies (below the zero value on the y-axis) and tragedies (above the zero value on the y-axis). The group assignment appearing over the ellipses thereby corresponds to the results of the KNN procedure. Once again, there is a clear intermediate area between comedies and tragedies along the zero value on the y-axis—at least from the viewpoint of the distributional semantics.

4. Discussion of the results and outlook

The comparison of the methods used here 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—significant results can be reached. Both methods are considered standard procedures in *text mining*. In order for the classification to reach a precision of over 70%, however, a comprehensive filtering of text was required, which, beyond punctuation and the usual stop words, made necessary the extraction of further function words, proper nouns and their nominal forms. The rest can only be manually assembled per body of work, which requires a lot of time and effort. A higher rate of classification precision can be reached considerably faster by, for one thing, conducting a massive reduction of the output matrix to a *sparsity* of 20%, and, for another thing, designating a manageably sized number of informative terms using the log-likelihood function, which, in turn, serves to provide a foundation for the classification of dramas through the k-nearest neighbor documents and a clustering based on the Euclidian distance. The results sought on the basis of 64 Calderón dramas could still be improved upon if all 110 of his written *comedias* were available for a computational evaluation. In the sense of the *digital humanities*, this conclusion represents an invitation to qualitative researchers to take a more accurate look at the texts they have already examined and to create characteristic word lists for each category in which typical terms for each can be distinguished. This particularly regards 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.

Interesting, too, are the intermediate developments introduced in this study, in which the results of the four explored methods are compared and on this basis six dramas of each category were identified which could be regarded with a high probability as being either tragedies or comedies. Here it is revealed that the Calderónian tragedies, obviously because of the way the words are used within the text, are much more reliably identifiable than the comedies. One example in particular, would be the historical-critical edition of »Amar después de la muerte«, introduced by Jorge Checa, a title unknown to the authors in this study before the analysis began, which illustrated that when compared to the research literature, its classification was subsequently verified. Because 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 the qualitatively working researchers to work systematically and to consistently implement these established criteria for classification on an entire sequence of plays. 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. It is, for instance, rather implausible that a title such as »Amor, honor y poder«, categorized among the *comedias* in the standard edition, and therefore not considered to be a tragedy, or in other words, a *drama* with a serious theme, as this play was consistently classified as a tragedy by all five of the methods applied here. Then again, in 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 *comedia mitologica*. The two dramas »La puente de Mantible« and »El castillo de Lindabridis« exhibit very strong tragedy signals in our analysis, whereas »Las fortunas de Andromeda y Perseo«, »Fieras afemina amor«, »El mayor encanto, amor« and »La fiera, el rayo y la piedra« exhibit strong comedy signals.[37] The status of this group of dramas-- as with those recognized by Parker and Sullivan as being »on the brink of tragedy«[38] – should therefore be discussed anew with regard to their designated categories. 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 processes points to this in an emphatic way.

Doubtless, the approach performed through distributional semantics contributes only one factor—albeit a very 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 110 *comedias nuevas* illustrates that the methods explored here provide qualitative researchers with

[37] The assessments of these works as “comedias mitologicas” by Castro de Moux 2001; Greer 1988; Cancelliere 2000, Arellano 2000, Peña-Pimentel 2011.

[38] Comp. with Parker 1988, p.58, 181, 182. Sullivan 2018, p.70, 316, 321.

indispensable information, which may stimulate further analyses. This, of course, would be more efficacious if all of the Calderónian dramas could be made available digitally and in a consistent publishing standard.[39]

5. 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, p.31–53.

Ignacio Arellano: Editar a Calderón. Hacia una edición crítica de las comedias completas. Frankfurt on the Main 2007.

Ignacio Arellano: Calderón y los géneros dramáticos, con otras cuestiones anejas. Honor, amor, legitimación política y autoridad de las taxonomías. DOI: 10.15581/008.34.1.100-126 In: Rilce. Revista de Filología Hispánica 34 (2018), Ed. 1, p.100–126. DOI: 10.15581/008.34.1.100-126

Walter Benjamin: Ursprung des deutschen Trauerspiels. Frankfurt on the Main 1978.

Piotr Bojanowski, Edouard Grave, Armand Joulin and Tomas Mikolov: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics 5 (2017), p.135–146.

John A. Bullinaria / Joseph P. Levy: Extracting semantic representations from word co-occurrence statistics. A computational study. DOI: 10.3758/BF03193020 In: Behavior Research Methods 39 (2007), p.510–526. DOI: 10.3758/BF03193020

Pedro Calderón de la Barca: Obras completas. Textos íntegros según las primeras ediciones y los manuscritos autógrafos. Ed. by Ángel Valbuena Prat / Luis Astrana Marín. Madrid 1951–1956.

Pedro Calderón de la Barca: Comedias y otras obras. Madrid 2007–2010.

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. DOI: 10.5209/tret.74021 In: Talía. Revista de estudios teatrales 3 (2021), p.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, p.55–76.

María Esther Castro de Moux: Alquimia y gnosticismo en Fortunas de Andrómeda y Perseo de Calderón: In: Actas del V Congreso Internacional de la Asociación: Internacional Siglo de Oro (AISO) Münster, 20-24 de julio de 1999. Ed. by Christoph Strosetzki. Frankfurt on the Main 2001, p.319–330.

Jorge Checa (Ed.): Pedro Calderón de la Barca: Amar después de la muerte. Edición y estudio. Kassel 2010.

Christophe Couderc: Le théâtre classique au Siècle d’or. Cristóbal de Virués, Lope de Vega, Calderón de la Barca. Paris 2012.

[39] 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.

Álvaro Cuéllar González: Stylometry and Spanish Golden Age Theatre: An Evaluation of Authorship Attribution in a Control Group of Undisputed Plays. Appears in: Digital Stylistics in Romance Studies and Beyond. Ed. by Christof Schöch / José Calvo Tello / Ulrike Henny-Krahmer / Robert Hesselbach / Daniel Schlör. Heidelberg 2022.

Hanno Ehrlicher: Einführung in die spanische Literatur und Kultur des Siglo de Oro. Berlin 2012.

Hanno Ehrlicher / Jörg Lehmann / Nils Reiter / Marcus Willand: La poética dramática desde una perspectiva cuantitativa: la obra de Calderón de la Barca. DOI: 10.5944/rhd.vol.5.2020.27716 In: Revista de Humanidades Digitales 5 (2020), p.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 / Frankfurt on the Main 2021.

Margaret Rich Greer: The play of power: Calderón's »Fieras afemina amor« and »La estatua de Prometeo«. In: Hispanic Review 56 (1988), Ed. 3, p.319–341.

Matthew Jockers: Macroanalysis. Digital Methods & Literary History. Urbana / Chicago / Springfield 2013.

Jörg Lehmann. Klassifikation von Tragödien und Komödien bei Calderón de la Barca. DARIAH-DE 2021. DOI: 10.20375/0000-000e-6729-1

Félix Lope de Vega: Arte nuevo de hacer comedias en este tiempo. Dirigido a la Academia de Madrid. Madrid: Alonso Martin 1621 [erstmals 1609]. [https://books.google.de/books?id=Ihh5oI6I4TsC].

Will Lowe: Towards a Theory of Semantic Space. Proceedings of the Annual Meeting of the Cognitive Science Society 2001, p.576–581.

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. Stuttgart 2003, P.285–327.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado und Jeffrey Dean: Distributed Representations of Words and Phrases and Their Compositionality. Proceedings of NeurIPS 2013, p.3111–3119.

Alexander A. Parker: The mind and art of Calderón. Essays on the Comedias. Cambridge / New York / New Rochelle 1988.

Miriam A. Peña-Pimentel: El Gracioso en el Teatro de Calderón. Un Análisis desde las Humanidades Digitales. Electronic Thesis and Dissertation Repository. 307. 2011. [https://ir.lib.uwo.ca/etd/3070]

Miriam A. Peña-Pimentel: Aplicación de mapas de tópicos al análisis semántico de algunas comediad de Calderón. In: Anuario calderoniano 5 (2012), p.115–130.

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. Kalamazoo 2018, p.99–125.

Sean Papay / Sebastian Padó / Ngoc Thang Vu: Addressing Low-Resource Scenarios with Character-aware Embeddings. Proceedings of the Second Workshop on Subword / Character Level Models. New Orleans 2018, p.32–37.

Yves Peirsman / Dirk Geeraerts / Dirk Speelman: The automatic identification of lexical variation between language varieties. DOI: 10.1017/S1351324910000161 In: Natural

Language Engineering 16 (2010), Ed. 4, p.469–491. DOI: 10.1017/S1351324910000161
Christof Schöch: Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama. In: Digital Humanities Quarterly 11 (2017), Ed. 2, p.1–53.
Christof Schöch: Fine-Tuning our Stylometric Tools. Investigating Authorship and Genre in French Classical Drama. In: Digital Humanities Conference 2013. Nebraska 2013.
Henry W. Sullivan: Calderón in deutschen und niederen Landen. Eine dreihundertjährige Rezeptionsgeschichte. Berlin 2017.
Henry W. Sullivan: Tragic Drama in the Golden Age of Spain. Kassel 2018.
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.
Peter D. Turney / Patrick Pantel: From Frequency to Meaning: Vector Space Models of Semantics. DOI: 10.1613/jair.2934 In: Journal of Artificial Intelligence Research 37 (2010), p.141–188. DOI: 10.1613/jair.2934
Ángel Valbuena Prat: Historia de la literatura española. Vol. II. 3. Aufl. Barcelona 1950, P. 479–571.
Joe H. Ward: Hierarchical Grouping to Optimize an Objective Function. In: Journal of the American Statistical Association 58 (1963), p.236–244.
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, p.177–195.

6. List of Images and legends

7. Attachment

Brief Description and Name of Drama	**Euclid Ward.D2**	**Euclid Ward.D2 sparse**	**POS Cosine**	**tf-idf Cosine**	**KNN**
T1-A secreto agravio, secreta venganza	C	T	T	C	
T2-El Alcalde de Zalamea	C	T	C	T	
T3-El magico prodigioso	T	T	T	T	
T4-El mayor monstruo del mundo	M	M	M	T	
T5-El medico de su honra	C	T	T	C	
T6-El pintor de su deshonra	C	T	C	C	
T7-El principe constante	T	T	M	T	
T8-La devocion 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 sueno	T	T	T	T	
T12-La-gran-Cenobia	T	T	M	T	
T13-Las tres justicias en una	C	T	T	C	
T14-Los cabellos de Absalon	T	T	T	T	
T15-Saber del bien y del mal	T	T	T	C	
C1-Casa con dos puertas mala es de guardar	C	C	C	C	
C2-También hay duelo en las damas	M	C	C	C	
C3-El encanto sin encanto	M	M	M	C	
C4-El Faetonte	M	M	M	M	
C5-El jardin de Falerina	M	M	M	T	
C6-El maestro de danzar	M	C	C	C	
C7-La dama duende	C	C	C	C	
C8-Los empeños de un acaso	M	C	C	C	
C9-Mejor esta que estaba	C	C	C	C	
C10-Peor esta que estaba	C	C	C	C	
C11-Primero soy yo	M	C	C	C	
C12-Mañanas de abril y mayo	M	C	C	C	
C13-Antes que todo es mi dama	M	C	C	C	
C14-No siempre lo peor es cierto	M	C	C	C	
C15-Dicha y desdicha del nombre	M	C	C	C	
Test1-Afectos de odio y amor.txt	M	M	M	T	T
Test2-El galan fantasma	C	C	C	C	C
Test3-Las fortunas de Andromeda y Perseo	M	M	M	M	C
Test4-Los dos amantes del cielo (T)	T	T	T	T	T
Test5-Amor, honor y poder (T)	T	T	T	T	T
Test6-La cisma de Ingalaterra (T)	T	T	T	T	T
Test7-En esta vida todo es verdad y todo mentira	M	M	M	M	T
Test8-La aurora en Copacabana	M	M	M	M	T
Test9-Las cadenas del demonio (T)	T	T	T	T	T
Test10-Amado y Aborrecido	M	M	M	M	T
Test11-Amar después de la muerte o el Tuzaní de la Alpujarra (T)	C	T	T	T	T
Test12-Las armas de la hermosura	M	M	M	T	T
Test13-Celos, aun del aire, matan	M	M	M	M	T
Test14-Darlo todo y no dar nada	M	M	M	T	T
Test15-Eco y Narciso	M	M	T	M	T
Test16-Fieras afemina amor	M	M	M	M	C
Test17-Luis Pérez el Gallego	C	T	C	C	C
Test18-El mayor encanto, amor	M	M	M	M	C
Test19-La púrpura de la rosa	M	M	M	M	C
Test20-El sitio de Breda (T)	T	T	T	T	T

Test21-Nadie Fíe Su Secreto (C)	C	T	C	C	T
Test22-No hay burlas con el amor (C)	M	C	C	C	C
Test23-El escondido y la tapada (C)	M	C	C	C	C
Test24-No hay cosa como callar (C)	M	C	C	C	C
Test25-Las manos blancas no ofenden (C)	M	M	M	C	T
Test26-Con quien vengo, vengo (C)	M	C	C	C	C
Test27-Céfalo y Pocris	T	T	T	T	T
Test28-La puente de Mantible	T	T	M	T	T
Test29-El castillo de Lindabridis	T	T	M	T	T
Test30-El monstruo de los jardines	M	M	M	M	C
Test31-La fiera, el rayo y la piedra	M	M	M	M	C
Test32-Para vencer a amor, querer vencerle	C	T	M	T	T
Test33-Lances de amor y fortuna	T	T	M	T	T
Test34-Hombre pobre todo es trazas	C	T	T	C	C