Classification of Genres through 500 Years of Spanish Literature in CORDE

José Calvo Tello

Abstract In this work I analyze the development of numerous genres in almost five hundred years of Spanish literature. For that, I use a large diachronic corpus composed by the *Real Academia Española*. After an introductory section, the dataset is described, focusing on some important aspects of the distribution and balance of the categories. Next, several classification tests are applied in order to find out which parameters lead to the highest scores, and what the results for each category are. The variance of these results is then explored using linear regression, resulting in the length of the text being a good predictor for the classification results. Finally, the historical evolution is analyzed, showing that the classification results for genres neither get better nor worse over time, but remain stable.

Keywords genre, classification, Spanish literature, corpora

1. Introduction

Several researchers have applied computational methods for genre analysis during the last 30 years. In general, works studying genres can be divided into two groups: the ones with a greater historical interest, looking closely at the development of a few categories (Raible 1980; Schrott 2015; Underwood 2016; 2019; Schröter 2019), and the ones analyzing a larger number of categories with a cross-sectional and comparative interest, trying to classify the different genres (Kessler, Numberg, and Schütze 1997; Stamatatos, Fakotakis, and Kokkinakis 2000; Santini 2011; Jockers 2013; Schöch 2013; Underwood 2014; Hettinger et al. 2016; Henny-Krahmer et al. 2018; Jannidis, Konle, and Leinen 2019). This difference with respect to the perspective of the research needs to be defined in an early stage of the work since it affects the data collection of texts and labels. Until now, only a few studies (like Biber and Conrad 2009) have tried to

Calvo Tello, José: Classification of Genres through 500 Years of Spanish Literature in CORDE, in: Hesselbach, Robert, et al. (Eds.): Digital Stylistics in Romance Studies and Beyond, Heidelberg: Heidelberg University Publishing, 2024, pp. 15–36. DOI: https://doi.org/10.17885/heiup.1157.c19362

combine both perspectives. However, the costs of this combination in terms of time and money easily exceeds the possibilities of many research projects: it requires collecting thousands of texts written in many centuries and labeled with dozens of categories.

Nevertheless, the collaboration between projects from different disciplines, such as digital humanities and corpus linguistics, can open up new possibilities. The access to data from institutions that have been gathering texts and metadata for years or even decades is especially interesting. For Spanish, the *Real Academia Española* (RAE) has played a central role in the composition of large corpora (Sánchez Sánchez and Domínguez Cintas 2007).

In this contribution, I combine the historical and the comparative perspective on genre analysis. The research question that I pursue is the following: how do various genres develop through the last five centuries of Spanish literature? Although certain works have looked closely at the development of very specific genres such as science fiction (Underwood 2016) or *historiette* (Raible 1980, 341), it is uncertain whether the status or influence of genres have changed over time (Todorov 1976). On the one hand, it is possible that genres had a greater influence in previous centuries in which the normative aspect of genres was perceived to be stronger. If this is the case, the classification results should deteriorate over time. On the other hand, the publishing sector has gone through an increasing professionalization starting in the nineteenth century that could have caused that texts belonging to a category becoming more clearly described in linguistic terms. If so, the classification results should improve over time. Of course, a third outcome is also possible: that the classification results of genres stay stable as time goes by.

In the next section, I will first describe the dataset used. Then, the results of the classification will be shown together with the influence of several classification parameters (algorithms, transformation, number and characteristics of tokens). In the third main section, several characteristics of the categories will be tested to explain the variance of the classification results. Finally, I classify the genres in each decade, looking at large historical patterns. To observe further details and the code used, the accompanying Jupyter Notebooks can be accessed online.[1]

2. Description of the Dataset: CORDE

The diachronic corpus CORDE (*Corpus Diacrónico del Español*) is one of the two corpora that the RAE launched in 2002, along with the contemporary *Corpus de Referencia del Español Actual* (CREA). Originally, the CREA was supposed to span the

1 https://github.com/cligs/projects2020/blob/master/stylisitcs_500_years_CORDE/code/Explanation%20of%20the%20results.ipynb

most recent 25 years, while CORDE would append every year the texts that became older than 25 years, considered already as historical (Sánchez Sánchez and Domínguez Cintas 2007). In this manner, every year would cause the modification of the boundaries of both corpora. However, the development of a new corpus, *Corpus del Español del Siglo XXI* (CORPES XXI), changed these plans. This new corpus covers the twenty-first century and adds new texts every year. For this reason, the development of the two original corpora of the RAE was frozen.

CORDE contains more than 34,000 texts and around 300 million tokens in its final version. Its material comes from all Spanish-speaking countries, although the majority are from Spain (74 percent of the texts). The corpus contains metadata about genres and topics, and literary works make up a considerable part of the corpus. The dates of production are between the eighth century (with only a few instances) and 1974 (Sánchez Sánchez and Domínguez Cintas 2007). The corpus aims to be a representative sample of the language for research purposes (Sánchez Sánchez and Domínguez Cintas 2007, 143; Rojo Sánchez 2010). Until now, the users have access to websites of CORDE or CREA that allow queries using several filters, such queries in the text, author, title of work, medium, etc. Its acceptance as a standard tool by Hispanic scholars is high (Kabatek and Pusch 2011), although it has also received critique by several researchers who have highlighted the poor philological quality of the medieval section (see a discussion in Rodríguez Molina and Octavio de Toledo y Huerta 2017).

Until recently, researchers could not have access to the full text version of these corpora. The reason given for this were copyright issues relating to the editions of the texts. However, in the last years the RAE has published more data extracted from the corpora and signaled its availability for research requests. In the case of my doctoral thesis (Calvo Tello 2021), I requested the frequencies of the tokens per document. This information cannot be protected by the status of the edition anymore since it does not contain the original text; it only contains facts about the analyzed object (more specifically, how frequently each token appears in each text). Along with this statistical data, several metadata fields were available as well.

To explore the historical development of the genres and topics I use the main label that the RAE assigns to each text. Table 1 shows the original labels and their translation into English.

To my knowledge, these labels do not follow any previous classification system or taxonomy of genres. Many of them could be grouped together in broader groups (such as verse, prose, expositive texts about a topic, etc.). However, it is necessary to recognize that finding a set of labels which can be used of a such a long span of time for such a wide community is an almost impossible task. These labels may be imperfect, but constitute a reasonable solution.

How is the distribution of these categories over time? Figure 1 shows the number of texts in each category over centuries.

Table 1 Labels in Spanish and English

Label in Spanish	Label in English
Artes y espectáculos	Arts and entertainment
Ciencias aplicadas	Applied sciences
Ciencias exactas, físicas y naturales	Exact, physical and natural sciences
Ciencias sociales y humanidades	Social sciences and humanities
Derecho	Law
Historia y documentos	History and documents
Prensa	Press
Prosa	Prose
Prosa didáctica	Didactic prose
Prosa dramática extensa	Extensive dramatic prose
Prosa narrativa breve	Short narrative prose
Prosa narrativa extensa	Extensive narrative prose
Religión	Religion
Sociedad	Society
Verso dramático breve	Short dramatic verse
Verso dramático extenso	Extended dramatic verse
Verso lírico culto	Cultured lyric verse
Verso lírico tradicional	Traditional lyric verse
Verso narrativo culto	Cultured narrative verse
Verso narrativo tradicional	Traditional narrative verse

The bars show a very irregular pattern: the eighth and ninth centuries have very few texts. From the tenth to the twelfth century the number of texts does not exceed one thousand instances. The rest of the centuries can be divided into two groups: centuries with more than 4,000 texts (the fifteenth, sixteenth, and twentieth century) and the rest, with nearly 3,000 instances. Neither the number of texts per century nor the distribution of genres and topics over centuries are balanced. In order to gain a theoretical perspective about corpora it could be desirable to obtain a similar number of texts for different variables such as genre or century (Schöch 2017), but this collides with the historic reality when looked at more closely. For example, it is not possible to balance drama in verse and prose over centuries: drama was written mainly either in verse or prose depending on the historical period. Likewise, it is unrealistic to expect journalistic texts during the Middle Ages or the sixteenth and seventeenth century. Only a few categories show a stable number of texts over more than three centuries and only one category offers a minimal level of stability over the 12 centuries: legal texts.

The number of texts in each category is one possible way of exploring the corpus. However, perhaps the categories have very different typical lengths in words. For

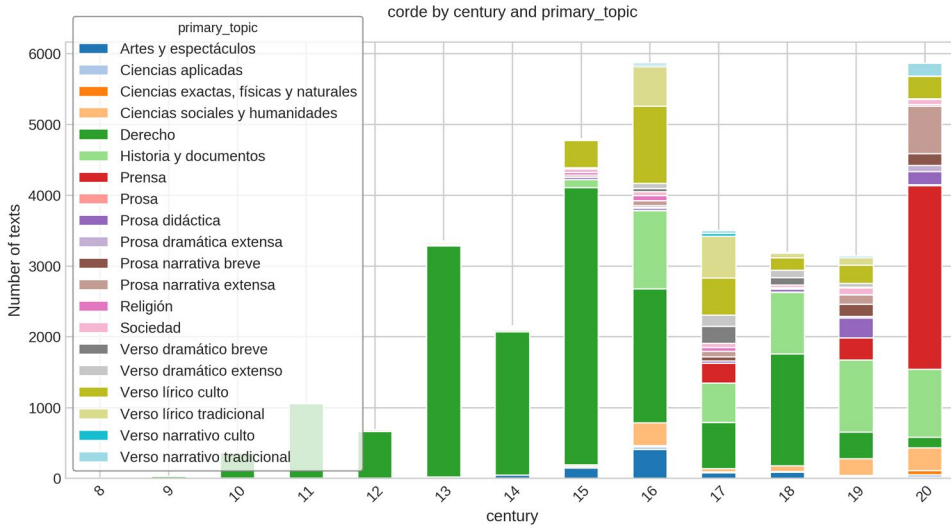

Fig. 1 Distribution of genres in number of texts by century (Calvo Tello, CC BY).

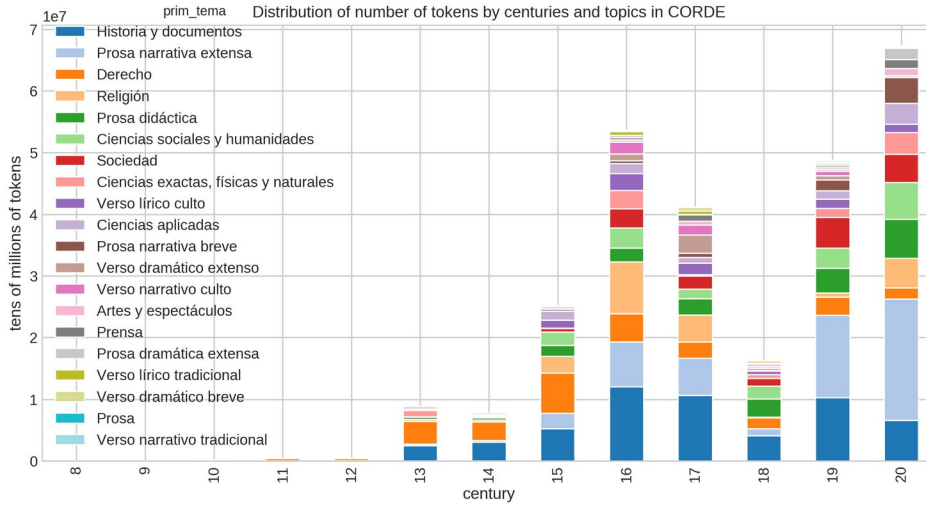

Fig. 2 Distribution of genres in number of tokens by century (Calvo Tello, CC BY).

example, journalistic texts are expected to be shorter than drama or prose fiction. For this reason, the Figure 2 shows the number of tokens populating the category in each century instead of the number of texts.

The overall picture does change notably. Until the twelfth century the pattern is mostly flat, increasing to around 10 million tokens in the thirteenth and fourteenth century. The rest of the centuries are populated by either around 20 million tokens (fifteenth and eighteenth century), or more than 40 million. Not only the total number of instances in each century has changed from Figure 1 to Figure 2, but also the distribution of each category. For example, legal texts predominate vastly in Figure 1, while this is much nuanced in Figure 2 when tokens are counted. A similar effect occurs with journalistic texts (red in Figure 1), whose proportion is reduced in Figure 2 (dark gray) because these texts are typically short. On the contrary, lengthy narrative prose (dominated by novels) has a larger proportion in Figure 2 than in Figure 1 because novels tend to be very long texts.

In any case, both bar plots show that the distribution of centuries and categories are not balanced, and the reason for that is to a certain point obvious: numerous categories do not exist in many of the centuries. Similar effects are observable when other variables of the corpus are taken into account, such as the country where the text was originated. Therefore, some countries have notably more texts in some centuries than in others, and some genres were more profusely written in some regions than in others (more details are to be found in the Jupyter Notebooks).

Instead of trying to artificially balance the corpus (Schöch 2017), I use it as it is and will use a subsampling approach in the analyses. In any case, several variables will be considered for explaining the classification results. However, the entire corpus is not considered: Due to the previous philological critique and the great deviation in terms of texts in the medieval section, only the texts from 1500 onward are analyzed. Additionally, many texts shorter than 100 tokens are also eliminated. With these restrictions, the remaining version of the CORDE contains 18,709 texts and over 192 million tokens. More details can be observed in the Jupyter Notebooks accompanying this publication.

3. Results and Parameters Evaluation

When genre classification is pursued, the researcher needs to define the variables of their data (language, categories, period) but also about the methods applied and which features exactly are accessed by the algorithm. Previous research has closely analyzed the effect of one or two of these so-called parameters (Kessler, Numberg, and Schütze 1997; Stamatatos, Fakotakis, and Kokkinakis 2000; Cerviño Beresi et al. 2004; Berninger, Kim, and Ross 2008; Santini 2011; Allison et al. 2011; Schöch 2013; Jockers 2013;

Underwood 2014; Hettinger et al. 2016; Calvo Tello 2019; Henny-Krahmer et al. 2018; Jannidis, Konle, and Leinen 2019). In this publication, I want to observe the effect of several methodological options. More specifically, I consider the following parameters:

— Categories: the 24 main categories defined by the CORDE
— Classifiers: ridge regression, support vector machines (SVC), logistic regression (LR), decision trees (DT), and random forest (RF)
— Transformation of the token frequencies: logarithmic transformation, z-scores, tf-idf, and binary
— Number of features: 100, 1,000, 2,000, 3,000, 4,000, 5,000
— Punctuation: kept or deleted

All these parameters have been tackled in the previous studies on text classification in the humanities or computer science cited above. How high the results of the classification will be for each possible combination of these five aspects is unknown. For instance, logistic regression when using 5,000 tokens without punctuation and transformed as tf-idf might be the best parameters for the classification of journalistic texts. For other categories, the highest results could be obtained with other parameters. These five aspects are tested in what computer science calls a *grid search* (Müller and Guido 2016, 262–64), that is trying every possible combination of the values of the parameters.

For the classification, a series of steps are taken to control the tests. First, for each category the corpus is split into two subcorpora of equal size, both containing the same number of texts. For example, the category *traditional narrative verse* contains 452 texts in the corpus. For its classification, the same number of texts from other categories are randomly sampled. These two subcorpora represent the instances belonging positively to the category, and those which do not. This produces a binary multi-label classification for each genre. This process is repeated five times to control the effect of the random samples of instances. The maximum number of instances in both subcorpora is settled to 1,000 texts to save computational costs. Finally, ten-fold cross validation is applied and the F1-score is measured in every test. A mean F1-score of this double loop is calculated, which constitutes here the data points of the following visualizations. The combination of all listed parameters above through the double loop produces a total of 276,000 iterations of classifications.

First, I would like to explore the results of the different categories annotated by CORDE. In Figure 3, the 50 highest scores for each category are summarized in box plots.

The red line at the bottom represents the baseline if an algorithm would ascribe to all instances the same category (0.50 since the corpus has been undersampled to create

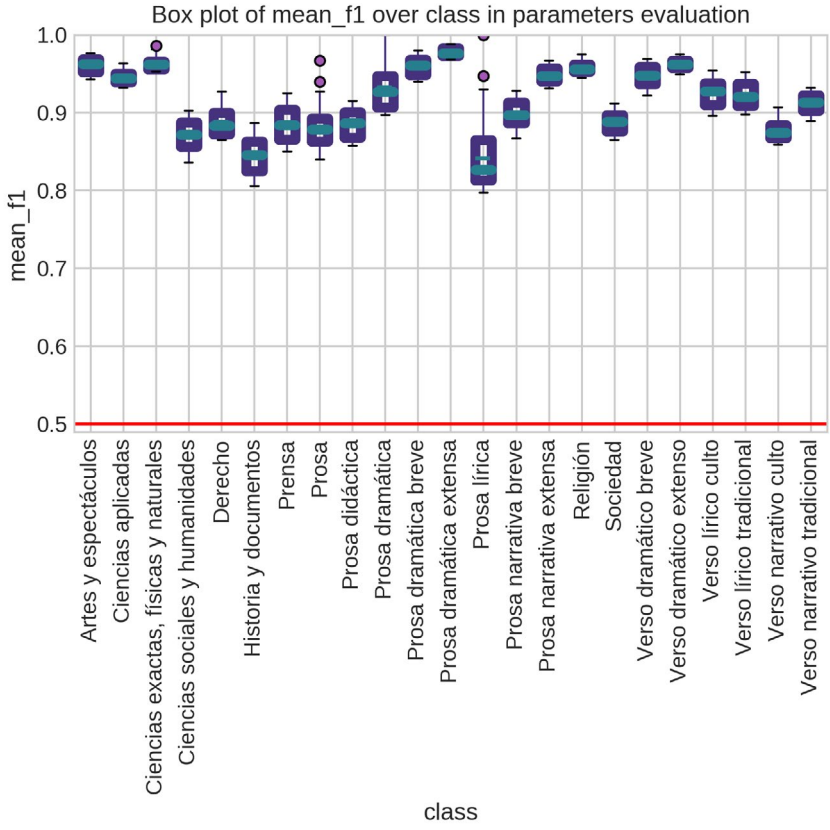

Fig. 3 Results of the classification process by categories (Calvo Tello, CC BY).

subcorpora of equal size). The results show that the classification results are clearly above this baseline. The median of the box plots are between 0.82 F1-score and 0.97 (*Artes y espectáculos, Ciencias exactas, Prosa dramática extensa*, or *Verso dramático extenso*) which are almost perfect results. The lowest scores are achieved by *Prosa lírica* (*Lyrical prose*), a hybrid category that can be easily accepted as difficult to differentiate from other categories. However, the next worst results are for *Historia y documentos* (*History and documents*), a category that I expected to show distinctive linguistic patterns that would differentiate it from the rest. Besides how high the results are, they also show certain variance, with some categories with all their 50 top results very close to each other (*Ciencias exactas, Prosa narrativa extensa*) while others encompass a variance of 10 percent (*Prosa, Prosa lírica, Prosa dramática*). This will be closely analyzed in the following section.

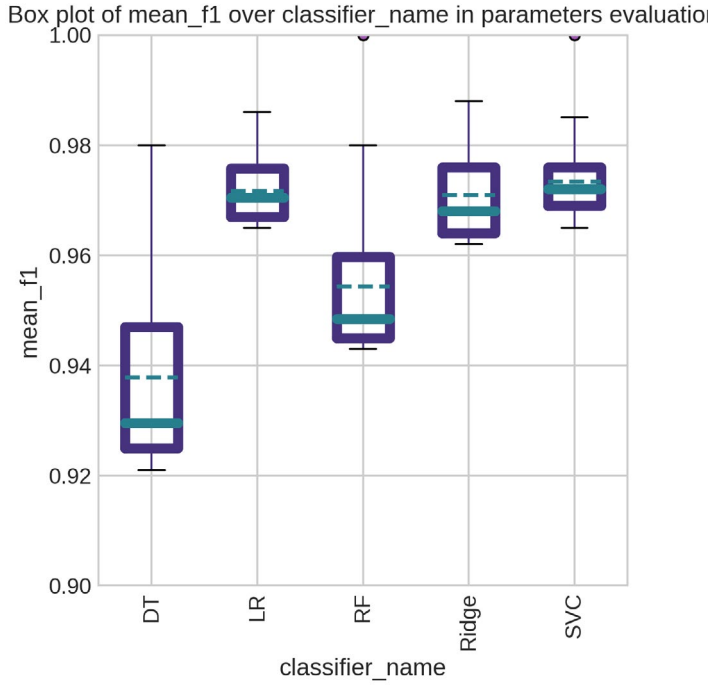

Fig. 4 Results of the classification process by classifiers (Calvo Tello, CC BY).

Later on, I show the top results of several parameters. Since the baseline of 0.5 F1-score is so clearly below the results, the next plots will not show it anymore. To maintain consistent figures, enable comparison, and gain in detail visibility of the results, the vertical axis will show from now on only the results between 0.90 and 1.00 F1-score. As in the previous figure, each box plot contains the top 50 results of each analyzed possibility. The results for the several classifiers are shown in Figure 4.

The box plots show two clear groups: decision trees and random forests with lower results, and logistic regression, ridge and support vector machines with similarly high results. A pairwise t-test between all these scores shows that the differences between these two groups are consistently significant ($p\text{-value} < 0.001$), while the differences between the algorithms with better performance are not significant (more details in the Jupyter Notebooks).

The second analyzed parameter is the transformation of the frequencies of tokens. The original data delivered by the RAE are first put in relation to the length of the document, obtaining the relative frequency of each token in each text. After that,

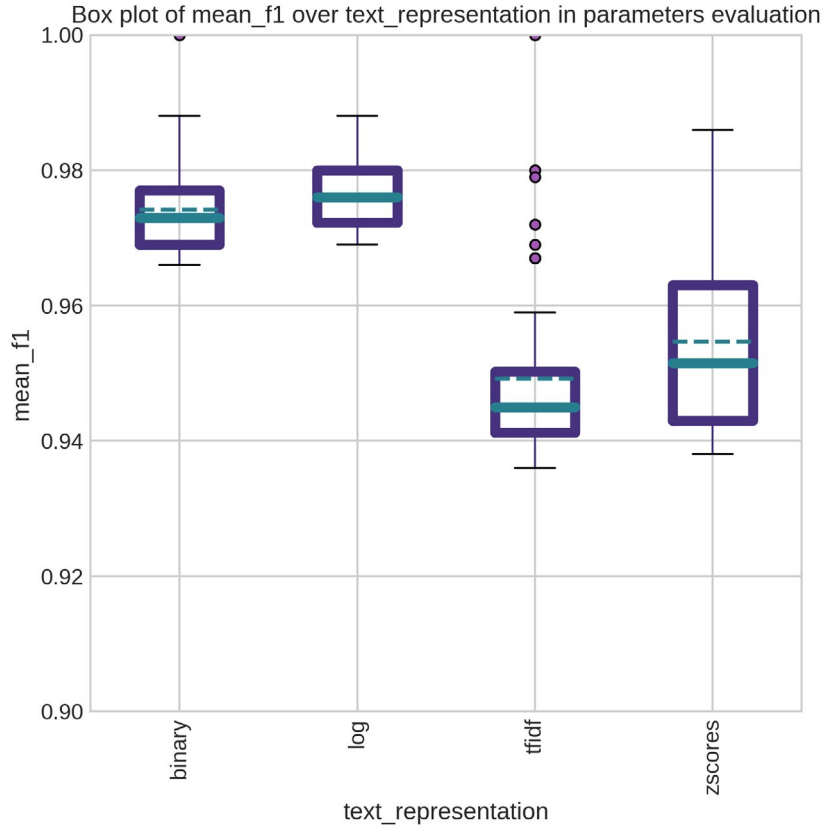

Fig. 5 Results of the classification process by transformations (Calvo Tello, CC BY).

these frequencies are typically transformed in other ways. In many cases this decision is associated with the field that the researcher feels closest to: z-scores in stylometry, tf-idf or binary frequency in computer science, or logarithm in statistics. Due to the corpus size, it was not feasible to run the algorithms using relative frequency.

Two groups of transformations can be observed in Figure 5: tf-idf and z-scores with lower results, and binary and logarithmic transformation with higher results. When these are compared pairwise, the two groups show statistical difference (p-value < 0.001), but the differences are not significant within the groups. In other words, logarithmic transformation does not yield statistically significant higher results than binary frequency (more details in the Jupyter Notebooks). Although these results are similar to what I have observed for my thesis (Calvo Tello 2021), binary frequency leads to

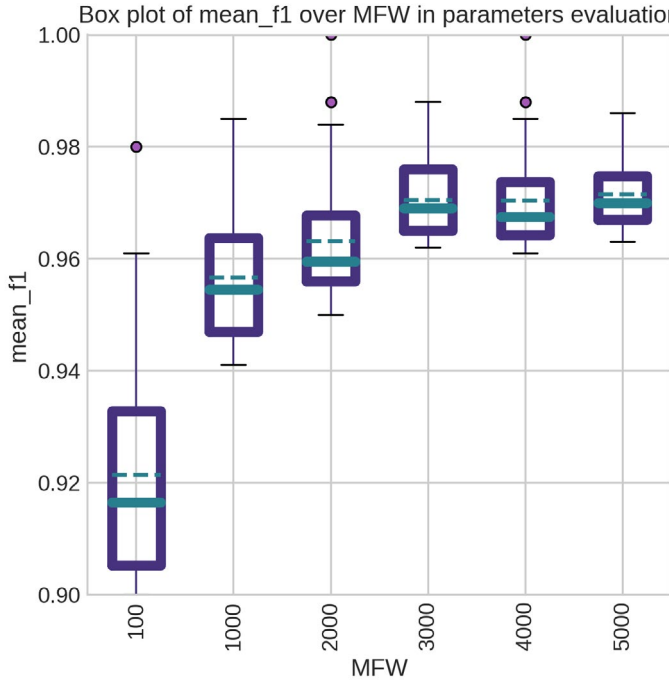

Fig. 6 Results of the classification process by number of features (Calvo Tello, CC BY).

notably higher scores in this case. This might be caused either by a greater number of texts or because the texts cover a much longer span of time. Further comparison in other corpora and languages would be of interest.

The next parameter is the number of features (frequently mentioned as most frequent words or MFW) passed to the algorithms. The results for this are shown in Figure 6.

The results improve up to 3,000 features (with statistical significance in every step, p-value < 0.01). From this point, the box plots overlap and there is no statistical improvement. These results are consistent with previous analyses that have also shown optimal results with around 3,000 features (Underwood 2014; Hettinger et al. 2016; Calvo Tello 2021).

The last analyzed parameter deals with the question whether the typographical tokens are good predictors for these categories or not. Although typographical characters have shown a positive effect in the classification results in other corpora (Calvo Tello 2021), in this case the results are nearly identical. The reason for this can be again

the fact that a long period is being analyzed. Even when the medieval section has been ignored for the analysis, the typographical conventions before the eighteenth century are notably different to the latter period. A lack of homogeneity in the philological transcription might diminish the positive effect of typography in this corpus. Further details about this parameter can be observed in the Jupyter Notebooks.

4. Explanation of the Classification Variance

Every publication about genre classification observes that there is a certain variance in the results: Some genres yield notably higher results than others. Although it might not strike anyone as particularly surprising that erotic novels can be classified with higher accuracy than social novels, the exact causes for that are still not fully explained. Previous research has put these differences in relation to the specificity of the features (Jockers 2013; Calvo Tello 2021), or to the disagreement in the labels of human institutions (Calvo Tello 2018). In this section, I want to observe whether the composition of the categories in the corpus might be influencing the results. For this purpose, I run a series of regression analyses comparing, for each category, two numerical variables: the F1-scores and a second variable that hypothetically should predict the classification results.

For example, a question that has been raised is whether genres constituted more robust textual categories in the past, but their importance has diminished over time (Todorov 1976). Especially for the period of modernity at the beginning of the twentieth century, several voices stated the decreasing influence of genres (Bradbury 1978; Calinescu 1987; Romero López 1997; Mainer, Alvar, and Navarro 1997; Buckley 2008; Longhurst 2008). To operationalize this hypothesis, I calculate the median year of publication of all the texts of each category. The median year of novels in the corpus is around 1900, while dramatic verse is typically earlier than 1700. These values are the independent variable, plotted in Figure 7 as the horizontal axis, while the vertical axis represents the mean accuracy of the classification results, measured in F1-scores.

The scatter plot shows that there is in fact a slightly decreasing tendency: The categories that were written in the earliest centuries of the analyzed corpus tend to yield higher scores than the ones published in latter ones. However, the results show a large deviation to the regression line. A regression analysis shows a p-value of 0.23 (in a Wald test, as the SciPy library applies for linear regression analysis). In other words: although there is a certain tendency to find lower classification results in latter centuries, it is not statistically significant.

One of the possible reasons for this is the fact that the median year of publication has reduced the thousands of data points of each publication into a single numerical value for each category. This reduction might have removed too much information.

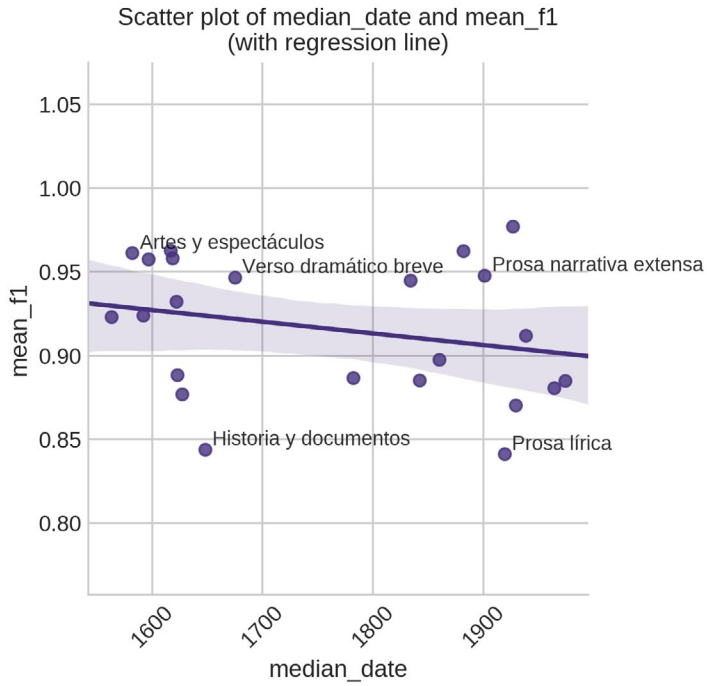

Fig. 7 Scatter plot with the categories by their median date and their classification results (Calvo Tello, CC BY).

Each category can also be captured by its chronological dispersion in the corpus. While journalistic texts are placed mainly in only two centuries, other categories cover the entire spectrum of centuries (*exact science, short narrative prose, legal texts*). To capture this, I have also calculated the standard deviation of the dates of publication, and use it as horizontal axis in Figure 8.

In this case the results show a much clearer tendency: The categories with greater standard deviation in their publication dates tend to show also higher classification results, like the category on the right of the scatter plot. On the contrary, *press* or *lyrical prose* show a narrower span of years of publications and this correlates with lower classification results. This tendency is statistically significant (p-value = 0.02).

However, this dispersion of the categories in the corpus might be hiding other lurking variables that correlate with the temporal dispersion. Perhaps the categories with a large standard deviation of years of publication are also the categories with more instances. More data per category might be beneficial to the classifier. To operationalize this, I measure the number of texts per category and apply it as an independent

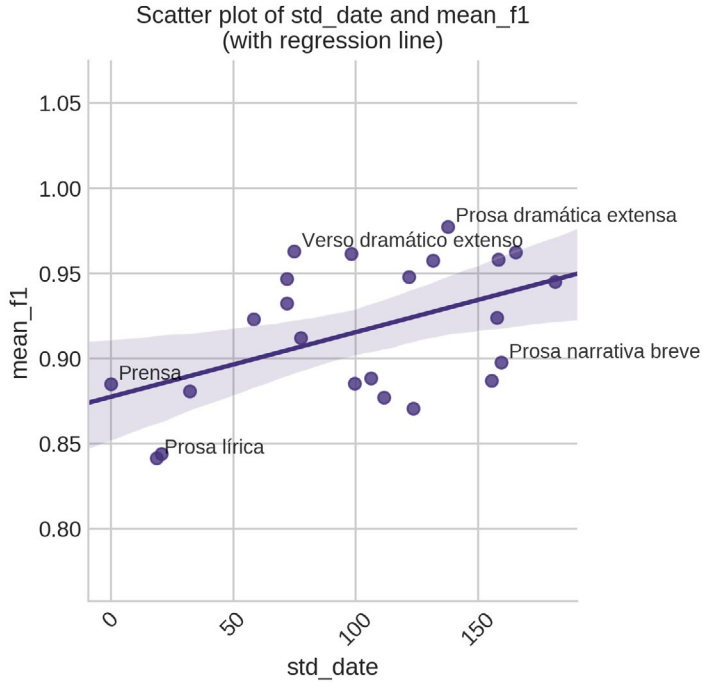

Fig. 8 Scatter plot with the categories by their standard deviation date and their classification results (Calvo Tello, CC BY).

variable. A regression analysis shows no correlation between both variables (p-value = 0.83). In this corpus, the size of the category cannot be used to predict the classification results (the scatter plot and further details can be observed in the Jupyter Notebooks).

The effect of a further variable is derived from Figure 3, which shows that the long versions of some categories tend to yield higher scores than the short ones. For example, *long narrative prose* achieves higher results than its short counterpart. This might be pointing toward a correlation between the length of the documents and the classification results: The more words a text has, the more data the classifier has to predict its category correctly. In the following figure, each category is represented by the median length in tokens of their texts. The position of the short and lengthy versions of several categories (*dramatic prose, dramatic verse, narrative prose*) can be read in labels of Figure 9.

It can be observed that in all cases, the long version of a category leads to higher results than the short one. For example *Prosa narrativa breve* (*short narrative prose*) has an F1-score around 0.90, while the *Prosa narrativa extensa* (*long narrative prose*) has an

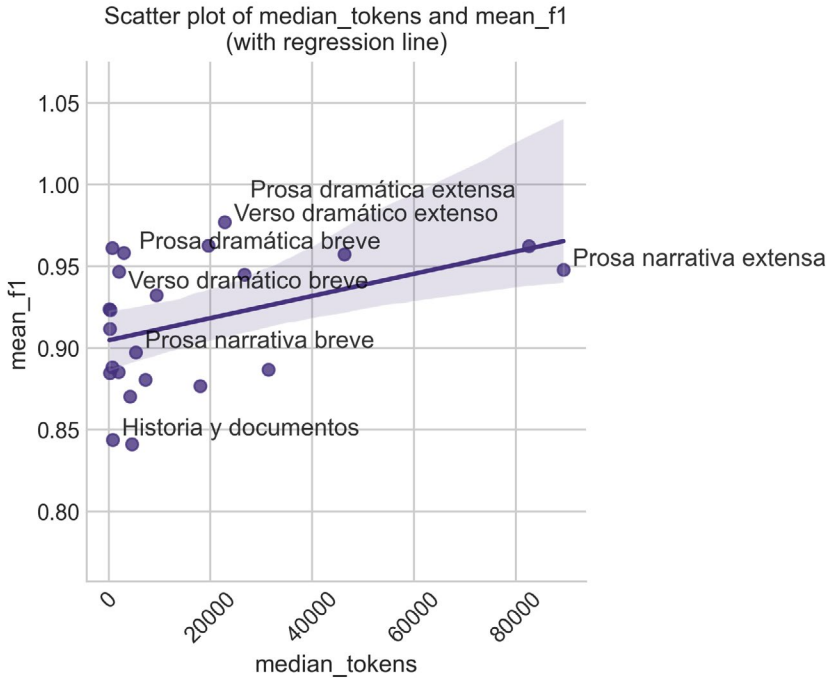

Fig. 9 Scatter plot with the categories by their median length in tokens and their classification results (Calvo Tello, CC BY).

F1-score of 0.95. A regression analysis gives a positive slope (p-value < 0.05), meaning that each additional token contributes to a slightly better classification result. This could also be the explanation for the low results of the category *Historia y documentos* (*history and documents)*: the texts pertaining to this category are among the shortest of the entire corpus.

5. Genre Classification by Decades: Historical Development of the Results

The most important question in this article is what is the general pattern of these categories over the five last centuries. Do the classification results become higher or lower over time? As mentioned earlier, there are good arguments for both possibilities.

Fig. 10 Scatter plot with the categories by their decade and their classification results (Calvo Tello, CC BY).

To observe the classification results from a historical perspective, I split the corpus in the 47 decades spanning between 1500 and 1970. In each decade I repeat the classification process explained in the previous section, analyzing all the categories. In Figure 10, each data point is one category in each decade.

The overall tendency is positive: the classification results tend to improve over time. This seems to support the second hypothesis that the genres have gone through a process of professionalization over time and therefore genres can be classified in later centuries more accurately using linguistic features. However, this improvement is not statistically significant (p-value = 0.066). In other words, genres can be classified better in later centuries, but that the hypothesis that this tendency is just produced by the random variability of the data cannot be rejected.

That is the general pattern, but how does each genre evolve over time? To answer this question, I have evaluated which categories do present a statistical tendency in their classification results. Only three of them show p-values under 0.05. Two of them improve their classification results: *Ciencias exactas, físicas y naturales* (*Exact,*

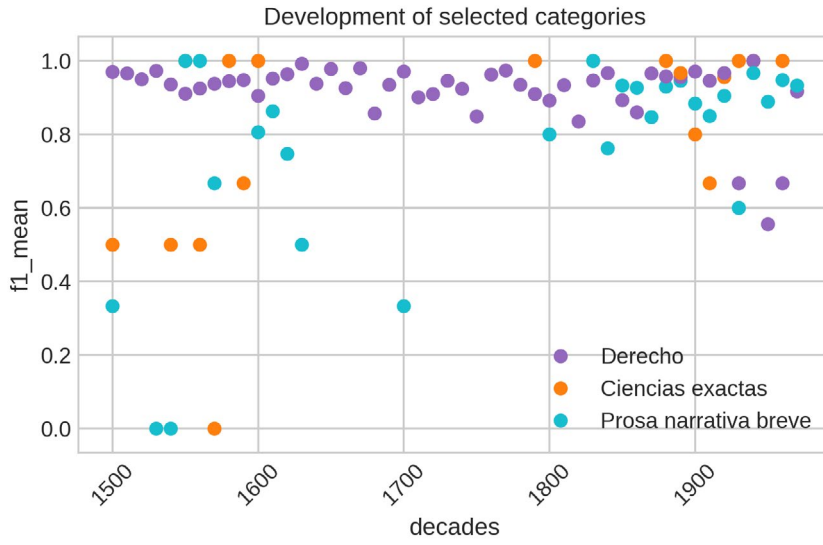

Fig. 11 Scatter plot with three categories for which the classification results show a statistical pattern (Calvo Tello, CC BY).

physical and natural sciences, p-value = 0.03) and *Prosa narrativa breve* (*Short narrative prose*, p-value = 0.002). In comparison, the classification of the category *Derecho* (*Law*, p-value = 0.007) tends to get lower results over time. The tendencies of these three categories are shown in Figure 11.

Besides these categories, the rest lead to very similar results over the entire period: Genres cannot be classified neither better nor worse as time goes by.

Figure 10 shows a positive tendency, although it is not statistically significant. Nevertheless, I want to explore whether this trend can be explained by spurious variables. In the previous section, I have observed that the length of the texts correlate with the classification results (Figure 9). The observed tendency in Figure 10 could be partially explained if texts become lengthier as time goes by. To explore this, Figure 12 represents each text as a data point, with its year of publication in the horizontal axis and its length in the vertical one.

Although there is a large variance in all periods, there is certain trend for longer texts in the later centuries than in the earlier ones. This can be evaluated through a regression analysis that confirms that texts tend to become longer, more specifically, 10 tokens longer each year on average (p-value < 0.001).

If the classification of these categories leads to higher results with longer texts and texts tend to become longer over time, the logical consequence is that classification

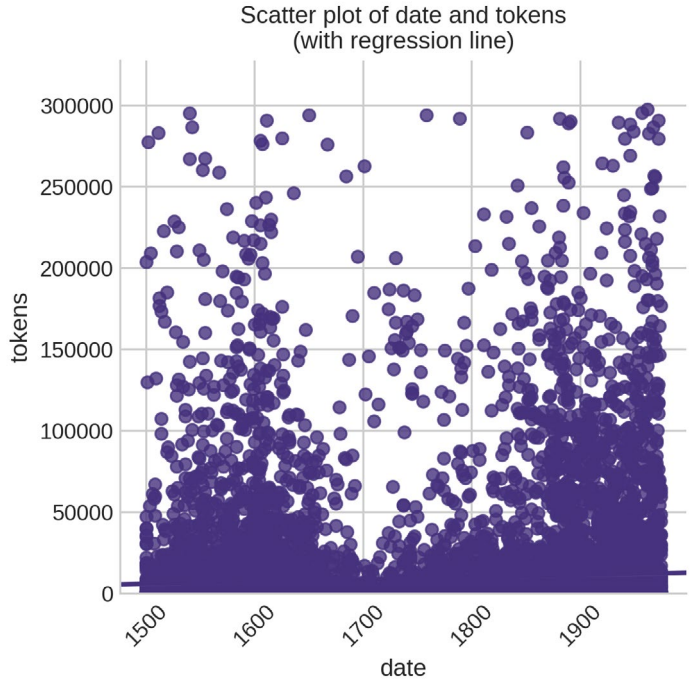

Fig. 12 Scatter plot with the texts by their year and their length in tokens (Calvo Tello, CC BY).

would tend to work better in the later centuries. This is what it was observed in Figure 10. To control for text length, I sample from the until now analyzed subcorpus, taking only the texts in the middle range of length. More specifically, I calculate the 25th and the 75th percentile of the tokens and retain only the texts within this middle range. That produces a corpus of 9,366 texts that are taken as the basis for a new classification test as described above.

The results of the classification in Figure 13 when the length of the texts is controlled shows exactly the opposite direction from the one observed in Figure 10. In this case, lower scores are obtained over time. However, as in the previous case, this tendency is not statistically significant (p-value = 0.10). The length of the texts seems to have an impact on the results: The direction of linear regression changes, and the p-value increases, meaning that the probability of the pattern being caused by randomness is higher. Overall, the results reject both hypotheses about the development of these categories over time: genres cannot be classified either better or worse over time; rather, the classification accuracy remains stable over time. This outcome contradicts

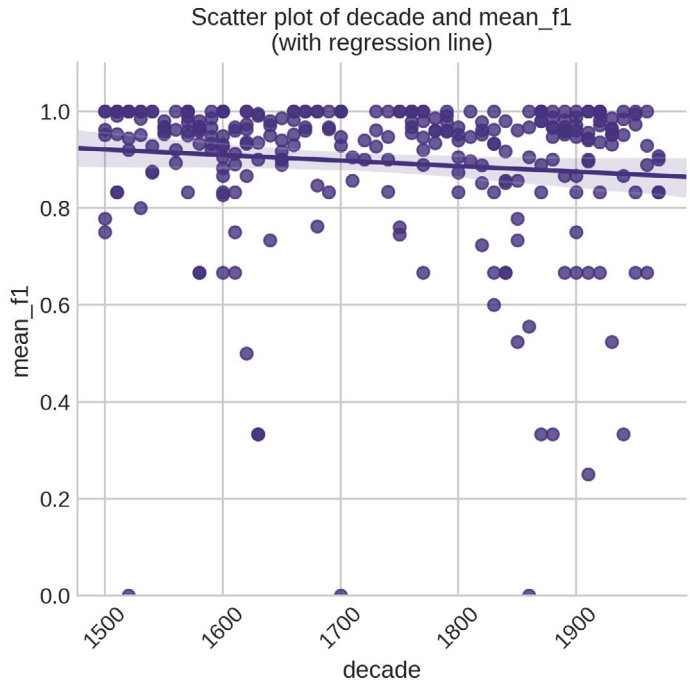

Fig. 13 Scatter plot with the categories controlling the length, by their decade and their classification results (Calvo Tello, CC BY).

the expectations from many literary scholars who, as cited before, assume that genres had greater influence, at least previous to the nineteenth century.

6. Conclusion

Have genres gained or lost importance over time? Did these labels describe the linguistic content of the texts better in earlier than in later centuries? To answer this leading question, I have used one of the largest diachronic corpora for Spanish composed by the RAE, which recently has opened up new possibilities of accessing the frequencies of tokens and metadata of their corpora. Larger datasets can reconcile the two perspectives on genre analysis: the cross-sectional and the historical one. Moreover, this research is an example about how corpora launched some decades ago can be currently combined with new methodologies such as machine learning.

The first main conclusion of this contribution is derived from the description of the dataset: The categories are neither balanced over time nor by other dimensions such as geographical origin. Although this can be seen as a defect in the composition of the corpus, a closer look reveals the impossibility of balancing genres such as dramatic texts (in prose or verse) or journalistic texts over periods of several centuries. The data show that some textual categories were produced more profusely during specific periods.

The second conclusion is that the multi-label classification yields very high results for all genres, some of them being very close to perfect scores. Moreover, the parameter analysis using grid search indicates that logistic regression (both in its classic version and as ridge regression) and support vector machines achieve better results than decision trees or random forest. Relating to the features, around 3,000 of them transformed logarithmically or binary tend to acquire the best results.

The third conclusion points toward two variables that can predict the classification results. In this corpus, the length of the texts of each category and the chronological span of the published texts can predict to a certain degree the classification results. In other words, if a genre tends to be populated by long texts, or if a genre was published over a long period of time, its classification results tend to be higher.

The final conclusion about the historical development shows that classification of genres remains stable over centuries. The results do not indicate either an increase or a decrease in the classification results. In other words, the majority of the genres neither gain nor lose linguistic distinctiveness. However, a closer look shows that two categories (technical scientific texts and short narrative prose) have a tendency to lead to higher classification scores in later centuries, while legal texts show the opposite trend.

This research offers a Jupyter Notebook companion which can be consulted to obtain deeper details than what is reported in this article.[2] Although these conclusions are relevant, it would be necessary to observe similar questions in several datasets. It would be of particular interest to analyze whether other sources of the labels of genres (information from the cover, literary scholars' opinions, etc.) lead to similarly stable patterns over centuries. Moreover, other languages might show dissimilar tendencies. Interdisciplinarity and collaboration between academic and national traditions can lead us to a deeper understanding of how cultures have changed over centuries through their texts.

ORCID®

José Calvo Tello https://orcid.org/0000-0002-1129-5604

2 https://github.com/cligs/projects2020/blob/master/stylisitcs_500_years_CORDE/code/Explanation%20of%20the%20results.ipynb

References

Allison, Sarah, Ryan Heuser, Matthew L. Jockers, Franco Moretti, and Michael Witmore. 2011. *Quantitative Formalism: An Experiment* (Stanford Literary Lab, Pamphlet 1). Stanford: Stanford Literary Lab.

Berninger, Vera, Yunhyong Kim, and Seamus Ross. 2008. "Building a Document Genre Corpus: A Profile of the KRYS I Corpus." In *BCS-IRSG Workshop on Corpus Profiling*. London: BCS Learning and Development Ltd.

Biber, Douglas, and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge, UK; New York: Cambridge University Press.

Bradbury, Malcolm, ed. 1978. *Modernism: 1890–1930*. Pelican Guides to European Literature. Hassocks: Harvester Pr.

Buckley, Ramón. 2008. "Tales from the Avant-Garde." In *A Companion to the Twentieth-Century Spanish Novel*, edited by Marta E. Altisent, 1. publ., 45–59. Woodbridge: Tamesis.

Calinescu, Matei. 1987. *Five Faces of Modernity: Modernism, Avant-Garde, Decadence, Kitsch, Postmodernism.* Durham: Duke University Press.

Calvo Tello, José. 2018. "Genre Classification in Spanish Novels: A Hard Task for Humans and Machines?" In *Data in Digital Humanities*. Galway: EADH.

Calvo Tello, José. 2019. "Gattungserkennung über 500 Jahre." In *DHd 2019 Digital Humanities: multimedial & multimodal. Konferenzabstracts*, 292–94. Frankfurt, Mainz: DHd.

Calvo Tello, José. 2021. *The Novel in the Spanish Silver Age: A Digital Analysis of Genre Using Machine Learning*. Bielefeld: transcript.

Cerviño Beresi, U., José Juan García Adeva, R. A. Calvo, and Hermenegildo Alejandro Ceccatto. 2004. "Automatic Classification of New Articles in Spanish." In *X Congreso Argentino de Ciencias de la Computación*, n. p. http://sedici.unlp.edu.ar/handle/10915/22551

Henny-Krahmer, Ulrike, Katrin Betz, Daniel Schlör, and Andreas Hotho. 2018. "Alternative Gattungstheorien: Das Prototypenmodell am Beispiel hispanoamerikanischer Romane." In *DHd 2018 Digital Humanities: Kritik der digitalen Vernunft. Konferenzabstracts*, 105–12. Köln: DHd.

Hettinger, Lena, Isabella Reger, Fotis Jannidis, and Andreas Hotho. 2016. "Classification of Literary Subgenres." In *DHd 2016 Digital Humanities. Konferenzabstracts*, 154–58. Leipzig: Universität Leipzig.

Jannidis, Fotis, Leonard Konle, and Peter Leinen. 2019. "Makroanalytische Untersuchung von Heftromanen." In *DHd 2019 Digital Humanities: multimedial & multimodal. Konferenzabstracts*, 167–73. Frankfurt am Main/Mainz: DHd. https://zenodo.org/record/2600812.

Jockers, Matthew L. 2013. *Macroanalysis – Digital Methods and Literary History*. Champaign, IL: University of Illinois Press.

Kabatek, Johannes, and Claus D. Pusch. 2011. *Spanische Sprachwissenschaft: eine Einführung.* Tübingen: Narr.

Kessler, Brett, Geoffrey Numberg, and Hinrich Schütze. 1997. "Automatic Detection of Text Genre." In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, 32–38. Stroudsburg, PA: Association for Computational Linguistics.

Longhurst, Carlos Alex. 2008. "The Early Twentieth-Century Novel." In *A Companion to the Twentieth-Century Spanish Novel*, edited by Marta E. Altisent, 30–44. Woodbridge: Tamesis.

Mainer, José-Carlos, Carlos Alvar, and Rosa Navarro. 1997. *Breve historia de la literatura española*. Madrid: Alianza.

Müller, Andreas C., and Sarah Guido. 2016. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. Beijing: O'Reilly.

Raible, Wolfgang. 1980. "Was sind Gattungen?" *Poetica* 12: 320–49.

Rodríguez Molina, Javier, and Álvaro Sebastián Octavio de Toledo y Huerta. 2017. "La imprescindible distinción entre texto y testimonio: el CORDE y los criterios de fiabilidad lingüística." *Scriptum digital: revista de corpus diacrònics i edició digital en llengües iberoromàniques* 6: 5–68.

Rojo Sánchez, Guillermo. 2010. "Sobre codificación y explotación de corpus textuales: Otra comparación del Corpus del español con el CORDE y el CREA." *Lingüística* 24: 11–50.

Romero López, María Dolores. 1997. "Hispanic Modernismo in the Context of European Symbolism – Towards a Comparative Dekon-Struction." *Orbis Litterarum* 52 (3): 194–210. https://doi.org/10.1111/j.1600-0730.1997.tb01978.x.

Sánchez Sánchez, Mercedes, and Carlos Domínguez Cintas. 2007. "El banco de datos de la RAE: CREA y CORDE." *Per Abbat: boletín filológico de actualización académica y didáctica* 2: 137–48.

Santini, Marina. 2011. *Automatic Identification of Genre in Web Pages: A New Perspective*. Saarbrücken: Lambert Academic Publishing.

Schöch, Christof. 2013. "Fine-Tuning Our Stylometric Tools: Investigating Authorship and Genre in French Classical Theater." In *Digital Humanities 2013: Conference Abstracts*. Lincoln: UNL: 383–6.

Schöch, Christof. 2017. "Quantitative Analyse." In *Digital Humanities: Eine Einführung*, edited by Fotis Jannidis, Hubertus Kohle, and Malte Rehbein, 279–98. Stuttgart: Metzler.

Schröter, Julian. 2019. "Gattungsgeschichte und ihr Gattungsbegriff am Beispiel der Novellen." *Journal of Literary Theory* 13 (2): 227–57.

Schrott, Angela. 2015. "Kategorien diskurstraditionellen Wissens als Grundlage einer kulturbezogenen Sprachwissenschaft." In *Diskurse, Texte, Traditionen: Modelle und Fachkulturen in der Diskussion*, edited by Franz Lebsanft and Angela Schrott, 115–46. Göttingen: V&R unipress.

Stamatatos, Efstathios, Nikos Fakotakis, and George Kokkinakis. 2000. "Automatic Text Categorization in Terms of Genre and Author." *Computational Linguistics* 26 (4): 471–95.

Todorov, Tzvetan. 1976. "The Origin of Genres." *New Literary History* 8 (1): 159–70.

Underwood, Ted. 2014. "Understanding Genre in a Collection of a Million Volumes, Interim Report." https://doi.org/10.6084/m9.figshare.1281251.v1.

Underwood, Ted. 2016. "The Life-Cycle of Genres." *Journal of Cultural Analytics* 2 (2). https://doi.org/10.22148/16.005.

Underwood, Ted. 2019. *Distant Horizons: Digital Evidence and Literary Change*. Chicago: The University of Chicago Press.