



FOR THE INDIGNITY OF THE TEXT:
MEASURING STYLOMETRIC ROBUSTNESS
IN AUTHORSHIP ATTRIBUTION
UNDER TEXTUAL CORRUPTION *

Álvaro CUÉLLAR
Universitat Autònoma de Barcelona (España)
alvaro.cuellar@uab.cat
<https://orcid.org/0000-0002-9934-6321>

Recibido: 24 de enero de 2026
Aceptado: 12 de febrero de 2026
<https://doi.org/10.14603/13E2026>

ABSTRACT:

The article addresses the dilemma between the philological demand for reliable texts in the digital environment and the practical need to work at scale with imperfect materials in Spanish Golden Age theatre. As a case study, it focuses on authorship attribution via stylometric methods and proposes a controlled experiment of progressive synthetic corruption on three corpora of secure authorship, in order to measure the degree of textual indignity a corpus can withstand before attribution performance deteriorates significantly. The methodology applies a standard classifier trained on relative frequencies of the most frequent words and evaluates performance as an increasing proportion of the text is replaced with external material. The results show substantial robustness across a wide degradation range and a gradual decline as corruption becomes dominant. The conclusion argues for methodological coexistence: preserving textual dignity as an editorial horizon while acknowledging that, under controlled conditions and with appropriate cautions, provisional texts can still support empirically grounded inferences of value to philology.

KEYWORDS:

Spanish Golden Age Theatre, Stylometry, Authorship Attribution, Textual Indignity, Synthetic Corruption.

* This research was supported by the projects ISTAE: Impresos sueltos del teatro antiguo español (II) (PID2022-136431NB-C66); La integral dramática de Lope de Vega (II) (PID2024-155554NB-I00); METADRAMA: Teatro áureo en diacronía: estudio, edición y puesta en escena del verso clásico (PID2024-161619NA-I00); Thal-IA: Patrimonio teatral áureo: Inteligencia Artificial y Fotografía Espectral (CNS2023-145014); and Tracing Regularities in Pedro Calderón de la Barca's Dramatic Oeuvre with a Computational Approach (DFG Project No. 508056339).

ARTENUEVO

Revista de Estudios Áureos

Número 13 (2026) / ISSN: 2297-2692

unhe
UNIVERSITÉ DE
NEUCHÂTEL

Institut de langues et
littératures hispaniques

POR LA INDIGNIDAD DEL TEXTO:
MEDICIÓN DE LA ROBUSTEZ ESTILOMÉTRICA
EN LA ATRIBUCIÓN DE AUTORÍA
BAJO CORRUPCIÓN TEXTUAL

RESUMEN:

El artículo aborda el dilema entre la exigencia filológica de textos fiables en el entorno digital y la necesidad práctica de trabajar a gran escala con materiales imperfectos en el teatro del Siglo de Oro. Como caso de estudio, propone la atribución de autoría mediante técnicas estilométricas y plantea un experimento controlado de corrupción sintética progresiva sobre tres corpus de autoría segura, con el fin de medir el grado de indignidad textual que puede soportar un corpus antes de que el rendimiento atribucional se deteriore de forma significativa. La metodología aplica un clasificador estándar entrenado con frecuencias relativas de las palabras más frecuentes y evalúa el rendimiento conforme aumenta el porcentaje de texto sustituido por material externo. Los resultados muestran una robustez notable durante un tramo amplio de degradación y un deterioro gradual cuando la corrupción se vuelve dominante. La conclusión defiende una convivencia metodológica: mantener la dignidad textual como horizonte editorial, pero reconocer que, bajo control y con cautelas, los textos provisionales pueden sostener inferencias empíricas de valor para la filología.

PALABRAS CLAVE:

Teatro del Siglo de Oro, estilometría, atribución de autoría, indignidad textual, corrupción sintética.



1. INTRODUCTION

Early modern Spanish drama constitutes a vast and invaluable literary archive—one that scholars have a responsibility to preserve, curate, and transmit. As philologists and historians, we act as custodians of cultural heritage, accountable not only for the survival of these texts but also for the conditions under which they circulate and are interpreted. This obligation now extends decisively to the digital environment, which has become indispensable for the preservation, dissemination, and accessibility of the Golden Age theatrical corpus. Yet digital mediation is not a neutral container: it introduces its own constraints, risks, and forms of degradation. For that reason, the digital dimension of textual stewardship cannot be treated as secondary or incidental; it demands sustained attention, methodological care, and an ethics of responsibility commensurate with the cultural value of the materials involved.

Valdés Gázquez (2024) articulates this problem with particular clarity in his study «Por la dignidad del texto. El teatro del Siglo de Oro y de Lope de Vega en la red. Principios ecdóticos y de Humanidades Digitales» («For the Dignity of the Text: Golden Age Theater and Lope de Vega Online. Editorial and Digital Humanities Principles»). His central claim is straightforward and difficult to contest: the migration of Golden Age theatre to the web has produced an abundance of accessible material, but a substantial portion of the processed texts circulating online lack the minimum philological solidity required for reliable reading or research. The issue is not merely that some digital versions are imperfect; it is that many are grounded in obsolete editorial criteria, uncollated witnesses, or unverified nineteenth- and early twentieth-century print traditions, in sharp contrast to the high standards achieved by modern critical editions in print. In this sense, the digital ecosystem has amplified a paradox: unprecedented availability coexists with systematic textual unreliability. Valdés Gázquez's argument is, above all, ecdotic and epistemological. The web's capacity for effortless duplication transforms ordinary editorial shortcomings into structural risks: once uploaded, a defective text does not simply contain errors—it propagates them. Because digital texts are copied, republished, repackaged, and reused across portals and projects, their corruptions can consolidate into an autonomous tradition of error. Worse still, those distortions can become effectively fossilized: for many users, the most visible online version is easily mistaken for the authoritative form of the work, acquiring the status of a stable

historical document by sheer repetition and discoverability. What looks like democratized access can therefore become an epistemic trap: scholarship and digital tools risk being built on unstable foundations, and interpretive claims risk inheriting the hidden biases and losses of degraded textual witnesses. The ethical implication follows: publishing in digital form does not lessen scholarly responsibility; it magnifies it. From this diagnosis Valdés Gázquez derives a programmatic defense of textual dignity in the digital sphere. Digital products must be treated with the same rigor we expect from print scholarship—not because the digital is inherently inferior, but because its scale and replicability intensify the consequences of philological negligence. His position does not deny the utility of quantity, nor does it reject open access; rather, it insists on transparent editorial accountability and on quality standards proportionate to use. At minimum, digital texts should clearly declare provenance and editorial method; ideally, the academic community should prioritize putting into open circulation the critical editions and reliable texts it already possesses, and do so using robust, interoperable frameworks (notably XML-TEI) and open-science principles. Where full digital critical editions are slow and resource-intensive, Valdés Gázquez also argues for pragmatic strategies—such as Minimal Computing approaches—that can accelerate dissemination without abandoning non-negotiable philological basics. The overall message is categorical: if we do not ensure that our digital textual artifacts are philologically reliable, we may end up harming the very heritage we claim to preserve—producing not a public good, but a scalable mechanism for the persistence and normalization of corruption.

And yet, precisely because this argument is right about the ethical and epistemic stakes of textual quality, the practical landscape forces a pragmatic reassessment. In many lines of research on Spanish Golden Age drama, textual indignity is not a desirable condition, but it is a recurrent and often unavoidable one. Depending on the aims of a project, we cannot always afford to wait until the relevant plays exist as fully verified critical editions—however ideal that situation would be. The scale of the theatrical archive makes this constraint structural rather than incidental. We are dealing with roughly three thousand comedias, and—being generous—perhaps only around a third can be said to be critically edited, thanks above all to the sustained editorial labour devoted to Lope de Vega and Calderón de la Barca over recent decades. If we insist that computational research must proceed only once the remaining two thirds are edited to modern critical standards, we are effectively accepting a delay of decades, if not longer. This tension leads to a dilemma that is at once methodological and institutional. Do we prefer to wait,

completing the puzzle piece by piece until the corpus is philologically complete, or do we allow intermediate textual states—provisional editions, rapid digital transcriptions, minimally curated texts—to serve a distinct, instrumental mission? Put differently: should we make room for minimal editions that are not adequate for every philological purpose, but are sufficient for specific analytical tasks, especially those that depend on scale rather than on the perfection of any individual witness? The question is not whether dignity matters—it does—but whether all scholarly questions require the same threshold of textual dignity to be answerable. A further complication is canonicity. If access to critically edited texts becomes the primary gatekeeping condition for large-scale analysis, research will continue to concentrate on the same safe authors and the same well-served segments of the repertoire—often for understandable reasons of feasibility, funding, and collaborative structure. The consequence, however, is a systematic distortion of our field of vision: the archive remains vast, but our analytical corpus remains narrow. In this context, the availability of imperfect texts is not simply a technical inconvenience; it is also a potential lever for expanding the empirical base of research, exploring understudied playwrights, and testing hypotheses that are otherwise confined to the already canonical subset. This is the space in which the present study intervenes: not to oppose a defence of philological rigor, but to ask—under controlled conditions—how far computational methods can operate responsibly when the textual substrate is, unavoidably, less than ideal.

A paradigmatic case is that of automatic transcriptions produced through OCR or HTR workflows. Their outputs are rarely perfect; they contain omissions, misreadings, segmentation problems, and a residue of noise that would be unacceptable as a basis for a critical edition. Yet it would be methodologically shortsighted to dismiss them wholesale as unusable. In many research scenarios, such transcriptions are not an endpoint but a provisional access layer—an intermediate representation that enables forms of inquiry that would otherwise remain impracticable. Even with a non-trivial error rate, automatically transcribed plays can sustain a broad range of scholarly operations. They can be indexed and searched, allowing researchers to locate motifs, names, formulae, or intertextual echoes that would be difficult to identify manually across large collections. They can be processed by downstream NLP pipelines to support exploratory analyses such as coarse thematic mapping, lexicon-based profiling, or other forms of large-scale descriptive work. They can also be routed through large language models to generate summaries, outlines, or descriptive reports that are often robust to local

noise and can help scholars triage materials, prioritize targets for closer inspection, or formulate hypotheses before investing in full philological work. The alternative is not a pristine edition; it is frequently no access at all. When the choice is between having no usable representation of thousands of early prints and manuscripts, or having an imperfect but searchable and computationally tractable transcription, the practical calculus is straightforward. This is not an argument against critical editing—on the contrary, critical editions remain the gold standard for interpretation, citation, and textual scholarship. It is, rather, an argument about sequencing and purpose: in a world where the ideal cannot be achieved everywhere at once, provisional transcriptions can function as bridges toward later editorial refinement while already enabling meaningful research in the present.

A second, closely related example concerns the digital reproduction of early modern documents. The ideal, of course, is high-quality imaging carried out according to the established protocols of libraries and archives: calibrated lighting, color targets, consistent resolution, stable metadata, and preservation-grade formats. Such standards are not cosmetic; they guarantee legibility, long-term usability, and scholarly accountability. Yet the gap between the ideal and the actual state of access remains substantial. For many collections—especially dispersed, underfunded, or only partially catalogued holdings—the pace of professional digitization is inevitably slow. In that context, a first-pass strategy can be epistemically valuable: a rapid, systematic capture of as many documents as possible using modest means (a basic camera or even a smartphone), producing images that may be imperfect but immediately consultable. The point is not to replace archival digitization, but to create an intermediate layer of accessibility that dramatically reduces the costs of discovery. Even lower-grade reproductions can support a range of fundamental scholarly tasks: identifying whether a witness is relevant at all; confirming the presence of a play, a paratext, or a specific scene; extracting rough readings for preliminary study; enabling remote collaboration; and guiding decisions about which items deserve subsequent, preservation-quality imaging. In many cases, such provisional access prevents research from being constrained by geography, funding cycles, or the long timelines of institutional workflows. The underlying question is again one of trade-offs and temporal horizons. Do we prefer to wait decades for comprehensive, fully standardized digitization to be completed—or can carefully acknowledged, deliberately provisional reproductions serve a legitimate scholarly function in the meantime? Framed this way, rapid but conscientiously imperfect digitization becomes not a betrayal of best practice, but a

pragmatic complement to it: a strategy that expands access now while leaving intact the long-term goal of producing durable, archival-quality digital surrogates.

In short, these approaches are not mutually exclusive, nor do they need to be framed as competing paradigms. High-standard, preservation-grade digitization remains the ideal endpoint and the necessary foundation for reliable scholarship; rapid, provisional capture functions as a pragmatic intermediate layer that accelerates access, discovery, and prioritization. Properly distinguished, transparently described, and used with methodological caution, both can coexist productively: one safeguards long-term integrity, while the other expands short-term availability and enables research to proceed without waiting for the entire archival and editorial infrastructure to reach completion.

In this article, we examine a concrete case that develops the argument outlined above: the use of stylometry for authorship attribution in Spanish Golden Age theatre. Stylometry, in its most established form, models authorial signal through the distribution of highly frequent words—function words and other common lexical items whose cumulative patterns tend to remain stable across works and genres. Under ideal circumstances, such analysis would rely exclusively on critically edited texts of maximal philological reliability. In practice, however, that ideal is often unattainable. The available materials frequently include automatic transcriptions, lightweight or provisional editions, and divergent witnesses whose textual differences—whether minor variants, omissions, or larger-scale distortions—are inseparable from the objects we wish to study. Here it is useful to separate two regimes that are often conflated. Valdés Gázquez argues from an ecdotic horizon in which the critical text is the scholarly end product, whereas the present study treats texts as measurement substrates for a delimited task (stylometric attribution). The point is therefore not to relax standards of textual dignity, but to quantify—under controlled conditions—the minimum textual reliability required for a stylometric inference to remain valid.

This raises a methodological question with immediate consequences for digital scholarship: should stylometric attribution be restricted to texts that have undergone full critical scrutiny, or can it operate meaningfully on imperfect textual data without forfeiting validity? Rather than answering this question impressionistically, we approach it experimentally. We test the resilience of stylometric attribution under controlled conditions by introducing progressive, synthetic corruption into a clean reference corpus. Through this artificial deturpation process, we quantify how much textual indignity the stylometric signal can tolerate before

attribution performance deteriorates—and, conversely, under what degrees and types of corruption stylometry remains operationally reliable.

2. EXPERIMENTAL DESIGN AND CORPORA

In contemporary practice, stylometry typically treats authorship not as an essence but as a probabilistic pattern emerging from recurrent linguistic habits rather than isolated expressive choices. Over the last decades, and particularly in recent work on early modern drama (Cerezo Soler y Calvo Tello, 2019; Cuéllar, 2024; Cuéllar y Vega García-Luengos, 2023a; Cuéllar y Vega García-Luengos, 2023b; Ferreira Barrocal, 2022; Hernández-Lorenzo y Byszuk, 2023; Lara Ramírez, 2025; Marcos Rodríguez, 2021; Martínez Carro, 2022; Ruiz Urbón, 2023a; Ruiz Urbón, 2023b; Vega García-Luengos, 2021), such frequency-based approaches have proved methodologically productive: they enable attribution and classification at a scale that would be impracticable under exclusively close-reading paradigms, and they offer a replicable way of testing hypotheses about authorial identity across large corpora. At the same time, the practical success of stylometric attribution has always sat uneasily with a philological concern that is especially acute in Golden Age studies: the textual object is rarely pristine (Eder, 2013). Even when we work with modern editions, plays often circulate in divergent witnesses and editorial states; and when we rely on digital corpora, we frequently depend on provisional transcriptions, automatic OCR/HTR outputs, or lightly curated texts whose error profiles are neither uniform nor fully documented (Camps, Clérice y Pinche, 2021). The methodological question, then, is not whether critical editions remain the gold standard for interpretation and citation; rather, it is how far stylometric classification depends on that standard (Eder y Rybicki, 2012; Eder, 2015; Eder, Rybicki y Kestemont, 2016). Put bluntly, to what extent can stylometric signal survive textual error, intervention, and corruption before attribution becomes unreliable?

To address this question empirically, we design a controlled degradation experiment using three corpora that are, in different ways, safe starting points. Each begins from texts whose authorship is secure, and each yields near-ceiling performance under a classical stylometric pipeline. All texts used in this study come from the ETSO project, which has been built thanks to the collaboration of numerous projects and individual contributors. ETSO aggregates, curates, and harmonizes materials from multiple origins and editorial states, and it is progressively making these texts available to the research community. The corpora employed here were

extracted from that common matrix, selecting only plays with stable, undisputed authorship for the experimental sets. The first corpus is a multiauthor reference set of one hundred plays attributed without dispute to nine playwrights: Lope de Vega, Calderón de la Barca, Tirso de Molina, Juan Ruiz de Alarcón, Francisco de Rojas Zorrilla, Agustín Moreto, Luis Vélez de Guevara, Antonio Mira de Amescua y Guillén de Castro (Cuéllar, 2024). Its role is diagnostic and foundational: it allows us to observe, in a balanced multiclass setting, how progressive corruption erodes classification when the only variable that changes is textual integrity. The second corpus is a large binary collection centred on Lope de Vega, consisting of 280 plays of secure Lope authorship and 749 plays by other dramatists of the period. The third corpus follows the same binary logic for Calderón, comprising 176 securely attributed plays by Calderón and 1,117 plays by other contemporaries. These two author-centred corpora serve a complementary purpose: they test robustness under high-volume conditions that better resemble the scale at which stylometry is often applied in practice, while keeping the attributional ground truth stable. Across all three corpora, the baseline classification setup is intentionally conservative and familiar to readers in the field. We employ a Support Vector Machine¹ trained on the z-scores² of relative frequencies for the 500 most frequent words³ (Pedregosa *et al.*, 2011). In their uncorrupted state, these corpora are classified with near-perfect accuracy: almost all plays are assigned to the correct category, whether in the nine-

¹ A Support Vector Machine (SVM) is a supervised classification method that learns from labelled examples, in this case plays of secure authorship, how to distinguish categories on the basis of quantitative features, here word-frequency profiles. It constructs a decision boundary that best separates one author or group of authors from others. We use a linear SVM as implemented in scikit-learn, a standard and well-established choice in stylometric attribution because it performs reliably when many linguistic features are involved.

² Z-scores standardize each frequency variable so that different word-features become directly comparable. For every word, its relative frequency in a given play is expressed in relation to the mean and standard deviation estimated from the training data. This scaling prevents very common words from dominating the model merely because of their magnitude and allows the classifier to focus on meaningful deviations from typical usage.

³ The most frequent words are used because they tend to reflect habitual linguistic practice rather than subject matter. In stylometry this set is largely composed of function words and other very common items, which are comparatively stable across topics and genres. Selecting 500 such words represents a methodological compromise: the number is large enough to capture a rich stylistic profile but limited enough to avoid instability associated with rare, content-driven vocabulary.

author multiclass scenario or in the binary settings centred on Lope and Calderón. This baseline matters for the logic of the experiment. Because the starting point is already close to the performance ceiling, any systematic loss in accuracy can be attributed to the progressive degradation imposed on the texts, rather than to an underpowered model or an unstable corpus. The central question is therefore straightforward, but methodologically non-trivial: how much can we degrade these corpora by introducing error and modification in a controlled, incremental fashion while stylometric attribution remains operationally reliable? By treating textual indignity not as an abstract worry but as an experimental parameter, the study aims to quantify the tolerance of frequency-based stylometry under conditions that mirror, in an explicit and measurable way, the imperfect textual realities of digital work on early modern drama.

A central methodological challenge in any controlled corruption experiment is not the choice of classifier but the construction of the corruption process itself. Computationally, it is difficult to generate noise that is genuinely irregular in the relevant statistical sense. Textual corruption in the transmission of Golden Age theatre can arise from many heterogeneous mechanisms: omissions, duplications, substitutions, contamination between witnesses, defective transcription, and a wide range of editorial interventions. Replicating such processes *in silico* is not straightforward, because rule-based procedures, even when parameterised probabilistically, tend to produce artefacts that are more regular than the historical phenomena they are intended to emulate. The major risk is therefore not insufficient degradation but excessive structure in the degradation model. Early versions of the present experiment attempted to simulate corruption by combining several operators through explicit rules: word duplication, word deletion, word substitution, and token-level degradation. These operators were applied with varying probabilities so that, for a given overall corruption rate, different proportions of each action would be triggered. In practice, however, these strategies repeatedly converged on a similar failure mode. The procedures produced a detectable deturbation signature: the altered texts began to resemble one another not because they shared an authorial profile, but because they shared the statistical texture introduced by the corruption algorithm. Under these conditions, the classifier could partially exploit regularities of the corruption process itself, and the experimental setting no longer functioned as a clean stress test of authorial signal. A subsequent approach sought to avoid this by replacing tokens rather than transforming them through internal rules. For each target work, a fixed percentage of its words was

substituted with words drawn from an external pool of texts. This removed some of the mechanical regularities of deletion and duplication, but it introduced a different form of stability. Because substitutions were performed at word level and distributed across the text, the injected material tended to converge towards the average lexical distribution of the external corpus. The result was again a relatively stable corruption profile, now shaped by the donor pool rather than by the target text.

For these reasons, we adopt in this study a deliberately simple, but methodologically more robust, strategy: block substitution. The aim is not to reproduce any single historical mechanism of corruption in all its philological detail. Rather, the aim is to introduce controlled amounts of exogenous material in a way that minimises learnable artefacts and maximises variability across realisations. We define the corruption level as a percentage of the target text. If the corruption level is 20 per cent, we replace roughly one fifth of the words of the play; if it is 50 per cent, we replace roughly half; and so on, up to the extreme case in which the entire text is substituted. The key design choice is that the replacement is not performed word by word across the whole play. Instead, we select one contiguous segment of the play whose length corresponds to the chosen percentage and replace that entire segment with a contiguous segment of the same length taken from another play. As the corruption percentage increases, the substituted block becomes progressively larger: at low levels it corresponds to a short passage, while at high levels it can approach the scale of the entire work. This approach has two properties that are crucial for the goals of the experiment. First, because each target text is paired with a different donor text and the donor segment is sampled at a random position, the injected signal is not stationary across the dataset. The altered texts therefore do not converge on a common corruption texture in the same way that rule-based noise often does. Second, the procedure models a meaningful form of textual indignity in digital environments: the presence of non-original material of heterogeneous provenance within a text that is treated as a single analytical object. The donor blocks are taken from a pool of approximately one thousand plays that are not used in the training or evaluation corpora by dozens of different authors. The donor pool is drawn from the same general dramatic domain as the experimental corpora and consists predominantly of comedias, with a smaller proportion of autos sacramentales and related forms available in ETSO at the time of extraction. Its generic profile is therefore broadly comparable to that of the main corpora, and the experiment does not rely on genre contrast as a source of variation. However, the donor pool

includes a higher proportion of automatic OCR/HTR transcriptions than the curated baseline corpora. As a result, the substituted blocks introduce not only exogenous authorial signal but also realistic transcription noise, including segmentation errors, misreadings, and orthographic instability. The corruption model thus approximates contamination within the same theatrical ecosystem, but under conditions that reflect the uneven quality typical of large-scale digital collections. Because block substitution involves random choices (the position of the replaced segment and the selection of the donor play and donor segment), each condition is instantiated multiple times. For every corruption level, we generate ten independent versions using different random seeds⁴. This repetition prevents the analysis from depending on idiosyncratic pairings between particular works, and it provides an empirical estimate of variability under stochastic corruption. In the results, this variability is reflected as a band around the mean performance curve.

The table below (Table 1) illustrates the logic of block substitution. From a 100-word excerpt of *La vida es sueño* (secure Calderón authorship), a contiguous block of 30 words has been replaced by 30 consecutive words taken from an automatic transcription of the comedia of unknown authorship *La bandolera de Baeza*; as a result, the substituted segment reproduces the orthographic inconsistencies, missing diacritics, and transcription errors characteristic of OCR/HTR output.

(...) que más cuidados le ofrece; sueña el pobre que padece su miseria y su pobreza; sueña el que a medrar empieza, sueña el que afana y pretende, sueña el que agravia y ofende; y en el mundo, en conclusión, todos sueñan lo que son, aunque ninguno lo entiende.	(...) que más cuidados le ofrece; sueña el pobre que padece su miseria y su pobreza; sueña el que a medrar empieza, sueña el que afana y pretende, detrás de unos muros altos donde no daba el reflejo de la luna pinto atisve parlando por su abujero
---	---

⁴ Because the corruption procedure involves random selection of donor texts and substituted segments, individual runs may vary. For each corruption level we therefore generate ten independent realisations and report the distribution of outcomes rather than a single score. Random seeds are fixed to ensure full reproducibility, and repetition allows us to estimate how much variability is attributable to chance rather than to the corruption level itself.

<p>Yo sueño que estoy aquí de estas prisiones cargado, y soñé que en otro estado más lisonjero me vi. ¿Qué es la vida? Un frenesí. ¿Qué es la vida? Una ilusión, una sombra, una ficción, y el mayor bien es pequeño; que toda la vida es sueño, y los sueños, sueños son. (...)</p>	<p>con su galán y en su estoque pinte pasados los cuerpos y soñé que en otro estado más lisonjero me vi. ¿Qué es la vida? Un frenesí. ¿Qué es la vida? Una ilusión, una sombra, una ficción, y el mayor bien es pequeño; que toda la vida es sueño, y los sueños, sueños son. (...)</p>
--	--

Table 1. Example of block substitution: original 100-word excerpt (left) and version with a 30-word contiguous replacement from an automatic transcription (right).

To separate the effects of corruption in the material being classified from corruption in the reference material used for learning, we evaluate three complementary scenarios. First, only the play being classified is deturpated, while the reference corpus used for training remains clean. This models a common research situation in which the target material is noisy, but the reference set is comparatively reliable. Second, the classifier is trained on a deturpated reference corpus, but it is asked to classify uncorrupted plays. This models the inverse situation: the available reference resources are of poor quality, while the object of analysis is well edited. Third, both the reference corpus and the target play are deturpated at the same corruption level. This corresponds to the situation in which all available textual materials are compromised, as often occurs when working at scale with heterogeneous digital collections. Classification performance is estimated through cross-validation. In the multiclass corpus of one hundred plays, we use a leave-one-out⁵ procedure so that each play is held out once and classified against models trained

⁵ Leave-one-out cross-validation evaluates performance when the dataset is relatively small. The model is trained on all plays except one and then tested on the excluded play; this process is repeated so that each play serves once as test material. The procedure maximizes the amount of training data in each iteration while ensuring that the evaluated play has not influenced the model.

on the remaining works. In the two binary corpora, we use 10-fold cross-validation⁶, so that in each fold a subset of plays is held out as test data while the remainder are used for training, and performance is averaged across folds. Results are summarised using the F1 score, reported as macro-F1⁷. This metric balances precision and recall and is less sensitive than raw accuracy to class imbalance, which is particularly relevant in the binary corpora where the target author constitutes a minority class. Across all settings, macro-F1 provides a consistent measure of how reliably the classifier assigns plays to their correct class as textual corruption increases. Taken together, these design decisions ensure that the experiment does not merely show that noise reduces performance, which would be unsurprising, but that it does so under a corruption model that is deliberately constructed to avoid stable, learnable artefacts. The purpose is to measure, under a realistic and variable form of textual indignity, how far stylometric attribution based on high-frequency lexical profiles remains operational before the injected material overwhelms the authorial signal.

3. EVALUATION OF RESULTS

Figures 1–3 report the same experiment under three corpus configurations: a nine-author multiclass setting (100 plays) and two large binary settings centred on Lope de Vega and Calderón, respectively. For each corpus, the plots show how macro-F1 changes as the proportion of substituted text increases, under three degradation scenarios: only the test text degraded, only the training corpus degraded, and both degraded. The shaded bands represent variability across ten independent

⁶ In 10-fold cross-validation the corpus is divided into ten subsets. In each iteration the model is trained on nine subsets and tested on the remaining one, and the process is repeated until every subset has been used as test data. The reported performance is the average across these runs, providing a stable estimate while maintaining strict separation between training and evaluation.

⁷ The F1 score combines two dimensions of classification quality: precision, which measures how often assigned labels are correct, and recall, which measures how successfully the model retrieves all items belonging to a class. Macro-F1 computes this balance independently for each class and then averages the results, giving equal weight to minority and majority categories. This is particularly important in the binary corpora, where the target author represents a smaller portion of the dataset and raw accuracy could otherwise be misleading.

realisations at each degradation level (10th–90th percentile), so each curve summarises not a single run but a distribution of outcomes generated by different random substitutions.

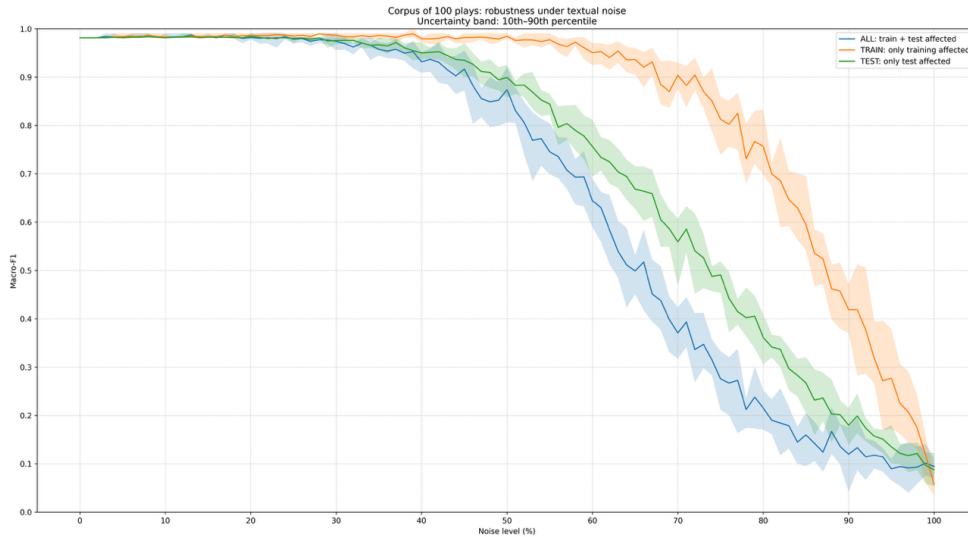


Figure 1. F1 score under progressive synthetic textual corruption in a nine-author attribution task (100 plays).

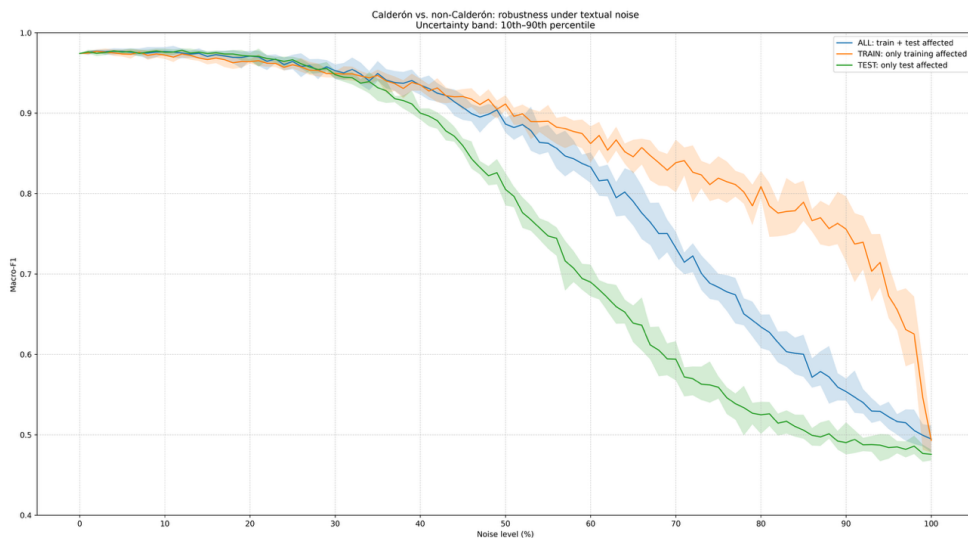


Figure 2. F1 score under progressive synthetic textual corruption in a Lope de Vega vs. non-Lope attribution task (280 Lope plays; 749 non-Lope plays).

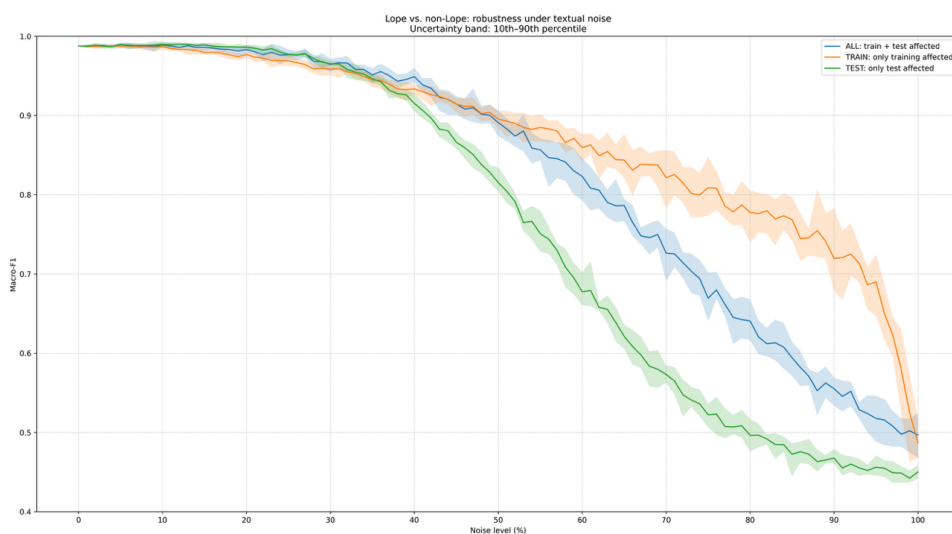


Figure 3. F1 score under progressive synthetic textual corruption in a Calderón de la Barca vs. non-Calderón attribution task (176 Calderón plays; 1,117 non-Calderón plays).

Before interpreting the substantive pattern, the figures provide an important validation of the experimental design. At zero degradation, performance starts near the ceiling in all three corpora, confirming that the corpora are internally coherent and that the stylometric pipeline is behaving as expected under clean conditions. This is not a minor detail. Because the baseline is almost perfect, subsequent declines can be interpreted as effects of progressive textual corruption rather than as artefacts of weak starting conditions. At the other extreme, as degradation approaches one hundred per cent, the curves converge toward the lower bound expected for arbitrary classification. In the multiclass case, random assignment among nine categories corresponds to a macro-F1 of approximately 1/9, that is, about 0.111. In the binary cases, random assignment corresponds to a macro-F1 of 1/2, that is, 0.5. The fact that the curves approach these task-specific limits is precisely what should happen if corruption eventually overwhelms authorial signal and if the procedure does not introduce a stable, learnable deturpation signature. This also explains why the y-axes differ between the multiclass and binary figures: the meaningful floor is not the same.

Within these validated endpoints, all three corpora display a remarkably consistent internal structure. First, there is a broad initial regime in which classification remains extremely high despite substantial textual alteration. In practical

terms, the system tolerates levels of deturpation that are far beyond the kinds of local instabilities that typically motivate philological concern in real attribution problems, such as sporadic corrupt readings, small editorial interventions, or limited transcription noise. Such issues usually affect only a very small fraction of the text, whereas in our experiment the model retains strong performance across degradation levels that correspond to a very large proportion of the play being replaced. The implication is not that textual quality is irrelevant for philology, but that the stylometric signal captured by high-frequency lexical profiles is highly redundant: it persists even when the textual surface is heavily disturbed. Second, across corpora the decline does not appear as an abrupt failure but as a progressive erosion. The curves remain flat or gently sloping through moderate degradation and then enter a clearer downward trajectory around the mid-range of the scale, where the substituted segment becomes a substantial portion of the text rather than a marginal disturbance. From that point onward, performance decreases steadily until it approaches the task-specific floor. In other words, the experiment does not suggest a sharp boundary between usable and unusable texts. It suggests a continuum with an extended region of operational robustness followed by a predictable degradation curve as the injected material begins to dominate the lexical profile. Third, the figures show a consistent asymmetry between the two one-sided degradation conditions. Corrupting only the test text is systematically more damaging than corrupting only the training corpus: performance drops earlier and more sharply when the text being attributed is deturpated than when only the reference material used for learning is degraded. This asymmetry is methodologically meaningful. It suggests that, in this stylometric setting, it is more harmful to degrade the particular text being attributed than to degrade the reference material on which the classifier is trained. A plausible interpretation is that when the test text is corrupted, the very object whose stylistic profile must be recognised is displaced; when only the training corpus is corrupted, the model is trained on noisier exemplars but is still asked to recognise a clean target, which appears to be a comparatively easier task. Put as a practical recommendation, if a research workflow allows selective improvement, it is especially valuable to ensure that the target text being attributed is as clean as possible, even when the surrounding reference corpora remain imperfect.

Taken together, the three figures support the same general conclusion. Frequency-based stylometry, implemented here with an SVM over z-scored relative frequencies of the 500 most frequent words, is robust to far more textual disturbance than is usually assumed in discussions that equate computational validity with

philological perfection. At the same time, the curves demonstrate that robustness is not unlimited: as the proportion of substituted material approaches and surpasses the mid-range of the text, performance begins to deteriorate systematically, and at extreme corruption it converges to the level expected for arbitrary classification. This combination of a near-perfect clean baseline, an extended robustness regime, and correct convergence to chance-level behaviour provides both substantive and methodological grounds for trusting the experiment's central claim: stylometric attribution remains operational under substantial textual indignity, but it eventually fails in a predictable way once degradation becomes the dominant component of the textual surface. These results have direct implications for authorship attribution in Spanish Golden Age theatre—they indicate that stylometric classification based on frequent-word profiles can remain reliable even under extreme textual disturbance: a play whose interior has been altered at levels that would be unacceptable for editorial purposes—for instance, around thirty per cent substitution in our corruption model—is still assigned correctly in most cases—this is methodologically encouraging for attribution work that must operate with provisional transcriptions, heterogeneous digital witnesses, or lightly curated corpora, because it suggests that a substantial margin of textual imperfection can be tolerated without collapsing the authorial signal. The present experiment is deliberately restricted to complete plays of secure single authorship. Collaborative works raise a different methodological problem. In such cases, attribution is typically performed at the level of individual jornadas rather than the play as a whole. Because each jornada contains substantially fewer words than a complete comedia, the available stylistic signal is correspondingly weaker and statistical estimates become more unstable. Under those conditions, any additional textual corruption would further reduce the recoverable authorial profile. The robustness threshold observed here for complete single-author plays should therefore not be transferred mechanically to collaborative drama without a dedicated analysis at the scale of individual acts. In this block-substitution setting, attribution remains reliably above chance across a wide range of corruption and begins to deteriorate systematically around the mid-range; as a pragmatic guideline, corruption levels on the order of 30% still preserve usable signal for single-author full plays in these corpora, while higher levels rapidly erode performance.

4. CONCLUSIONS

This article began from a philological imperative. The digital circulation of Golden Age drama demands an ethic of responsibility: if our digital artefacts are not philologically reliable, they do not merely contain error, they reproduce it at scale. Yet the same digital environment that magnifies corruption also makes visible a practical constraint that is structural rather than incidental. If large-scale research were to wait for the completion of critical editing across the entire theatrical archive, many empirical questions would be postponed for decades and the analytical corpus would remain confined to the best-served, most canonical segments of the repertoire. Between these two poles lies the space in which textual indignity becomes both a risk and, under certain conditions, a resource. Rather than treating indignity only as a normative category, the study approaches it as an empirical variable: something that can be increased gradually, measured, and observed in relation to analytical performance. To do so, it implements a controlled degradation experiment in which three corpora with secure authorship and near-ceiling baseline performance are progressively deturpated. The multiclass corpus of one hundred plays attributed to nine authors provides a stringent diagnostic setting, while two large binary corpora, centred on Lope de Vega and Calderón, test robustness under conditions closer to the scale at which stylometry is often applied in practice. Across all three corpora, the analysis relies on a standard stylometric pipeline: an SVM classifier trained on z-scored relative frequencies of the 500 most frequent words, evaluated with cross-validation and summarised using macro-F1.

The methodological problem, however, is not simply how to classify, but how to corrupt. Computationally, generating textual degradation that is both realistic and sufficiently irregular is difficult. Rule-based noise tends to introduce unintended regularities, and word-level substitution can drift toward the average lexical profile of the donor pool, producing a stable deturpation signature. In response, the experiment adopts a deliberately simple but robust strategy: block substitution. A contiguous segment of each play, sized according to the chosen corruption level, is replaced with a contiguous segment drawn from an external pool of roughly one thousand plays, most of them produced through automatic transcription workflows. This design injects not only exogenous authorial signal but also heterogeneous OCR/HTR artefacts, approximating the mixed and uneven conditions under which large digital corpora often circulate. Repeating the procedure ten times per corruption level allows the results to be interpreted not as a single

trajectory but as a stable tendency under many possible random substitutions. Across the three corpora, the behaviour of the curves supports a coherent interpretation. Stylometric attribution remains highly stable across a substantial initial range of deturpation, far beyond what would correspond to the kinds of small local uncertainties that often preoccupy editors and historians in practical attribution work. Performance then declines progressively as the substituted material becomes a large fraction of the text, and eventually reaches the expected regime in which the authorial profile is no longer recoverable. The comparison between scenarios also yields a consistent practical insight: degrading the text being classified is more damaging than degrading the training material, suggesting that, when selective improvement is possible, priority should be given to making the target text as clean as feasible, even when the available reference corpora remain imperfect.

What, then, becomes of indignity? These findings do not license complacency about error, nor do they reduce the long-term necessity of critical editing for interpretation, citation, and cultural transmission. The point is different. In computational practice, the relationship between philological perfection and analytical usefulness is not binary. There exists a substantial middle ground in which imperfect texts, explicitly acknowledged as such and used with methodological caution, can support reliable stylometric inference. In that sense, indignity can be productive: not because error is desirable, but because provisional textual states can function as access layers that enlarge the empirical horizon of research, reduce canonic bias, and enable questions that would otherwise remain untestable at scale. A constructive future for Golden Age studies in the digital sphere depends on keeping these commitments together rather than placing them in competition. Philology and computation should not be cast as rival epistemologies, but as complementary forms of care for the archive, operating on different timescales and with different deliverables. One task is to build and maintain islands of verified reliability that anchor interpretation and provide calibration points for digital methods. The other is to make responsible use of abundant, heterogeneous, and provisional corpora so that scholarship can proceed at scale without pretending that scale is neutral. The broader message is therefore one of methodological coexistence: to defend textual dignity as a long-term horizon, while also recognising that controlled, well-understood forms of textual indignity can be a legitimate, and in some contexts necessary, instrument for research in the present.

WORKS CITED

- CAMPS, Jean-Baptiste, Thibault CLÉRICE y Ariane Pinche, «Noisy medieval data, from digitized manuscript to stylometric analysis: Evaluating Paul Meyer's hagiographic hypothesis», *Digital Scholarship in the Humanities*, 36 (Supplement 2), 2021, págs. 49–71.
- CEREZO SOLER, Juan y José CALVO TELLO, «Autoría y estilo. Una atribución cervantina desde las humanidades digitales. El caso de La conquista de Jerusalén», *Anales Cervantinos*, 51, 2019, págs. 231–250.
- CUÉLLAR, Álvaro, «Stylometry and Spanish Golden Age Theatre: An Evaluation of Authorship Attribution in a Control Group of One Hundred Undisputed Plays», en *Digital Stylistics in Romance Studies and Beyond*, ed. de Robert Hesselbach, José Calvo Tello, Ulrike Henny-Krahmer, Christof Schöch y Daniel Schlör, Heidelberg, Heidelberg University Publishing, 2024, págs. 101–117.
- CUÉLLAR, Álvaro y Germán VEGA GARCÍA-LUENGOS, «La francesa Laura». El hallazgo de una nueva comedia del Lope de Vega último, *Anuario Lope de Vega. Texto, literatura, cultura*, 29, 2023a, págs. 131–198.
- , «Un nuevo repertorio dramático para Andrés de Claramonte», *Hipogrifo. Revista de literatura y cultura del Siglo de Oro*, 11.1, 2023b, págs. 117–172.
- EDER, Maciej, «Mind your corpus: systematic errors in authorship attribution», *Literary and Linguistic Computing*, 28.4, 2013, págs. 603–614.
- , «Does size matter? Authorship attribution, small samples, big problem», *Digital Scholarship in the Humanities*, 30.2, 2015, págs. 167–182.
- EDER, Maciej y Jan RYBICKI, «Do birds of a feather really flock together, or how to choose training samples for authorship attribution», *Literary and Linguistic Computing*, 28.2, 2012, págs. 229–236.
- EDER, Maciej, Jan RYBICKI y Mike KESTEMONT, «Stylometry with R: A Package for Computational Text Analysis», *The R Journal*, 8.1, 2016, págs. 107–121.
- FERREIRA BARROCAL, Jorge, «Estudio de una comedia del Siglo de Oro atribuida a Adrián Guerrero: *El ignorante discreto*», *Studia Aurea*, 16, 2022, págs. 283–307.
- HERNÁNDEZ-LORENZO, Laura y Joanna BYSZUK, «Challenging Stylometry: The Authorship of the Baroque Play *La Segunda Celestina*», *Digital Scholarship in the Humanities*, 38.2, 2023, págs. 544–558.

- LARA RAMÍREZ, Alberto, «Propuesta de un análisis estilométrico para las comedias colaboradas de Rojas», *eHumanista: Journal of Iberian Studies*, 61, 2025, págs. 345–352.
- MARCOS RODRÍGUEZ, Emma María, «Texto, atribución y censura de Próspera y Adversa fortuna de Don Álvaro de Luna», *Talía: Revista de estudios teatrales*, 3, 2021, págs. 79-89.
- MARTÍNEZ CARRO, Elena, «“David perseguido y montes de Gelboé”. Novedades estilométricas ante una atribución dudosa», *Anuario Lope de Vega. Texto, literatura, cultura*, 28, 2022, págs. 401–422.
- PEDREGOSA, Fabian *et al.*, «Scikit-learn: Machine Learning in Python», *Journal of Machine Learning Research*, 12, 2011, págs. 2825–2830.
- RUIZ URBÓN, Cristina, «Nuevas pruebas estilométricas para negar la autoría cervantina del Entremés de los romances», *Studia Aurea*, 17, 2023a, págs. 545–583.
- , «Sobre la validez de los análisis cuantitativos en los estudios de autoría de textos breves: el caso particular de los entremeses del Siglo de Oro», *Ogigia: Revista electrónica de estudios hispánicos*, 33, 2023b, págs. 69–96.
- VALDÉS GÁZQUEZ, Ramón, «Por la dignidad del texto. El teatro del Siglo de Oro y de Lope de Vega en la red. Principios ecdóticos y de Humanidades Digitales», en *Editar el Siglo de Oro en la era digital*, ed. de Susanna Allés-Torrent y Eugenia Fosalba, Barcelona, Studia Aurea Monográfica, 2024, págs. 71-108.
- VEGA GARCÍA-LUENGOS, Germán, «Las comedias de Lope de Vega: confirmaciones de autoría y nuevas atribuciones desde la estilometría (I)», *Talía. Revista de estudios teatrales*, 3, 2021, págs. 91–108.