ANUARIO Lope de Vega texto · literatura · cultura

THE MODERNIſA PROJECT: ORTHOGRAPHIC MODERNIZATION OF SPANISH GOLDEN AGE DRAMAS WITH LANGUAGE MODELS

Javier de la Rosa (Nasjonalbiblioteket i Norge),
Álvaro Cuéllar (Universitat Autònoma de Barcelona)
y Jörg Lehmann (Staatsbibliothek zu Berlin)

Cita recomendada: Javier de la Rosa, Álvaro Cuéllar y Jörg Lehmann, «The Moderniſa Project: Orthographic Modernization of Spanish Golden Age Dramas with Language Models», *Anuario Lope de Vega. Texto, Literatura, Cultura*, XXX (2024), pp. 410-425.

DOI: <https://doi.org/10.5565/rev/anuariolopedevega.530>

Fecha de recepción: 12 de octubre de 2023 / Fecha de aceptación: 3 de noviembre de 2023

Resumen

La creciente aplicación de métodos computacionales a la literatura española del Siglo de Oro ha revelado la necesidad de automatizar la modernización de los textos para facilitar su comparación y análisis. Este estudio es el primero en el uso de técnicas del Procesamiento del Lenguaje Natural (PNL) para adaptar los textos del Siglo de Oro (*ca*. 1590-1680) a un español moderno y normalizado (RAE 2010). La investigación emplea la arquitectura de transformadores para entrenar y evaluar modelos usando un corpus de comedias del Siglo de Oro. Dichos modelos son prometedores a la hora de encargarse de marcas tipográficas complicadas, así como palabras dependientes del contexto, pero se ven comprometidos al tratar los nombres propios y las variaciones ortográficas. Evaluada usando diferentes métricas comunes en la literatura especializada, nuestra herramienta demuestra tener potencial como recurso valioso para historiadores, filólogos y humanistas digitales. Las limitaciones incluyen la especificidad del corpus de entrenamiento y algunas inconsistencias observadas en la puntuación y la ortografía incluso en textos modernizados. Esta investigación ofrece una solución novedosa y escalable a la modernización manual de la literatura del Siglo de Oro, abriendo la puerta a más estudios computacionales en el ámbito de conocimiento.

Palabras clave: Transformadores; modernización automática; ortografía; inteligencia artificial; Humanidades Digitales; Siglo de Oro.

Abstract
The increasing application of computational methods to the literature of the Spanish Golden Age has revealed the necessity of automating the modernization of its texts to facilitate seamless comparison and analysis. This study pioneers the employment of Natural Language Processing (NLP) techniques for the transformation of Spanish Golden Age texts (circa 1590-1680) into modern, normalized Spanish (RAE 2010). The research employs the transformer architecture to train and evaluate models using a corpus of Golden Age dramas. The models show promise in handling tricky typographical marks and context-sensitive words, but also struggle with proper nouns and orthographic variations. Evaluated using different metrics common in the specialized literature, the tool demonstrates potential as a valuable resource for historians, philologists, and digital humanists. Limitations include the specificity of the training corpus and observed inconsistencies in punctuation and spelling even in modernized texts. This research offers a novel, scalable solution to the manual modernization of Golden Age Spanish literature, enabling further computational studies in the field.

Keywords: Transformers; Automatic Modernization; Orthography; Artificial Intelligence; Digital Humanities; Spanish Golden Age.

Introduction

The application of computational analysis to Spanish literature, and to the Golden Age period (16th-17th centuries) in particular, has grown in interest in recent years (de la Rosa and Suárez 2016, Cerezo Soler and Calvo Tello 2019, Demattè 2019, Fiore 2020, García-Reidy 2019, Vega García-Luengos 2021 and 2023, Cuéllar 2023, Cuéllar and Vega García-Luengos 2023). For most of this research (*e.g.,* stylometry, sentiment analysis, automatic dating), digital editions in a modern and homogenized orthography are usually preferred (Cuéllar and Vega García-Luengos 2017-2023, Vega García-Luengos 2023). Most digitization pipelines apply automatic recognition (OCR or HTR) to identify the characters of a text as printed, and traditional philologists transcribe texts as faithfully to the original as possible. While new approaches try to improve the existing OCR systems to produce modernized text directly (Cuéllar 2021a and 2021b), the vast amount of readily available digitized material in digital libraries and archives cannot be easily re-processed. In addition, there is a genuine interest in modernization among historians and literature editors, who would benefit greatly from automatic modernization. Unfortunately, we failed

to find such systems for Spanish,[1] although several historical language models exist for other languages and purposes (Manjacavas and Fonteyn 2021, 2022a and 2022b, Schweter *et al.* 2022, Gabay *et al*. 2022). In this work, we demonstrate how techniques from natural language processing (NLP) can be employed to transform Spanish texts available with historical orthography (*ca.* 1590-1680) into modern normalized Spanish (RAE 2010). In providing a pre-trained model usable for the whole community of philologists, historians, digital humanists, and editors, we hope to foster research with regard to the given timeframe and to establish an alternative to the current cumbersome approach of transcribing Golden Age texts manually.

Methodology

The development of the transformer architecture (Vaswani *et al.* 2017) caused a paradigm shift in NLP. Transformer-based language models excel at many tasks from coherent narrative generation to question answering, and from any sort of classification task to translation (Brown *et al*. 2020, He *et al*. 2021, Liu *et al*. 2020, Xue *et al*. 2021). Alas, creating these models requires billions of words, thousands of hours of computation, and many tons of carbon emissions dropped into the atmosphere (Strubell *et al.* 2019). The bright side is that once a pre-trained language model (PLM) exists, it can be adjusted (fine-tuned) to a specific downstream task with limited data in a fraction of the time and the resources. In this work, we approach orthographic modernization as a translation task and fine-tune existing language models on a parallel corpus of Spanish Golden Age dramas. The majority of PLMs work with vocabularies that might split words into smaller sub-word units called tokens (Devlin *et al.* 2019). The more frequent a word appears in the pre-training corpus, the higher the probability of keeping the word intact. Since orthographic modernization is a character-based process, we tested both token-free and token-based PLMs. In particular, we fine-tuned the multilingual versions of text-to-text transformers T5 (mT5) and ByT5 (Xue *et al.* 2021 and 2022) for the trans-

1. Normalization alternatives exist as part of multilingual toolkits that deal with OCR post-correction (*e.g.,* Reynaert *et al.* 2015).

lation of 17th-Century Spanish to modern Spanish. We then evaluated the results using the BLEU metric (Papineni *et al.* 2002). In order to avoid misinterpretations of the translation metric caused by the similarity between 17th-Century Spanish and Modern Spanish (Post 2018), we complemented the metric with the average word and character error rates (WER and CER). As our baseline, we calculated all metrics using the historical orthography with no changes against the modernized version of the same texts.

Corpus Construction

We built a parallel corpus of Spanish Golden Age theater texts with pairs of Golden Age orthography and current orthography. For the old orthography, we used the Teatro Español del Siglo de Oro (TESO) corpus,[2] because there each text is «copied exactly as it is written, with all peculiarities captured: accents, abbreviations, etc.» (TESO Editorial Policy, online).[3] For the current orthography, we used the Corpus de Estilometría aplicada al Teatro del Siglo de Oro (CETSO), a collection of modern editions of the same and many more texts (Cuéllar and Vega García-Luengos 2017-2023). We chose 44 dramas by the Golden Age dramatists Juan Ruiz de Alarcón, Pedro Calderón de la Barca, Félix Lope de Vega Carpio, and Juan Pérez de Montalbán. All dramas were published in Madrid and Barcelona between 1614 and 1691 for the first time and were written in verses of similar metrical characteristics. Both corpora were aligned line by line to establish a ground truth for the translation between the different varieties of Spanish.

2. Online, <https://quod.lib.umich.edu/t/teso/>. Accessed on 5th September 2023. The original texts of the TESO database are in the public domain; the protection rights of the database expired in Europe in the year 2013, 15 years after the publication of the database. Because the database has been established and published in Europe in 1998, the US-American company ProQuest LLC, which purchased the database, cannot claim a copyright for it, because it is subject to European regulations. Compare <https://quod.lib.umich.edu/t/teso/CofU.html>. Accessed on 5th September 2023.

3. It is important to acknowledge that despite this statement, we are aware that the TESO corpus represents a characteristic orthography that ultimately adheres to specific norms and forms of transcription. Consequently, it is plausible that other transcriptions may exhibit variations in their presentation of orthography. Future developments could augment this work by incorporating additional orthographic variants, thereby offering a more nuanced corpus for the process.

Results

After randomizing all 141,023 lines in the corpus, we split it into training (80%), validation (10%) and test (10%) sets stratifying by play. We then fine-tuned mT5 and ByT5 base models on sequence lengths of 256 doing a grid search for 3 and 5 epochs, weight decay 0 and 0.01, learning rates of 0.001 and 0.0001, and with and without a "translate" prompt. Table 1 shows the results on the test set of the best model on the validation set for each model type.

Model	**BLEU**	**WER**	**CER**
Baseline (no changes)	48.04	32.19%	8.95%
mT5	79.22	14.96%	4.48%
ByT5	**80.66**	**14.17%**	**4.20%**

Table 1. Scores for baseline and the best models on the test set. Best scores in bold.

While both models perform modernization reasonably well, ByT5 seems to be outperforming baseline and T5. We applied our best model to an unseen play (*Castelvines y Monteses* by Lope de Vega, 1647) and analyzed the errors produced. We discovered that the model is capable of solving some difficult corner cases in typographical marks (*e.g.,* adding initial exclamation marks) and some other tricky words («*cómo*» vs «*como*», «*qué*» vs «*que*») by leveraging contextual information. However, it struggles with proper nouns that normally would go uppercase (*e.g.*, «Castelvines», «Monteses»). We also discovered some strange artifacts in our ground truth corpus regarding archaisms and homogeneity of spelling that might have impacted the learning of the models (*e.g.,* «*efeto*» should appear as «*efecto*» 'effect', «*agora*» as «*ahora*» 'now').

Figure 1 shows an example of the modernization result of *Castelvines y Monteses* by Lope de Vega, where we can appreciate the correctness of the process. Here we can observe the first verses of the play and the transformation between the original text with historical orthography (left) and the result of our process of modernization (right). The overall results seem satisfactory (errors in red, correct transformations in green). Most words are correctly modernized, except for two quite

complicated situations: the infrequent usage of «*Ardese*» in «*Ardese la casa toda*», that should have been modernized as «*Árdese la casa toda*» and the syntactically strange utilization of the pronoun «*Tu*» in «*Tu lindo gallina*», that should have been modernized as «*Tú, lindo gallina*». In this last case, the lack of the comma in the original text confused the process.

Ardese la casa toda	**Ardese** la casa toda
de fiesta, y de regozijo.	de fiesta y de **regocijo**.
Casa, alguna hija, o hijo?	Casa, ¿alguna hija o hijo?
O es el concierto, o la boda.	O es el concierto o la boda.
Ve por tu vida Marin,	Ve, por tu vida, **Marín**,
y entra al descuydo.	y entra al **descuido**.
Harto bien,	Harto bien,
porque en colacion me den	porque en **colación** me den
las exequias de mi fin,	las exequias de mi fin,
en casa de tus enemigos,	en casa de tus enemigos,
me mandas entrar a ver?	me mandas entrar a ver?
Pues quien te ha de conocer?	Pues, **¿quién** te ha de conocer?
Para mal, siempre ay testigos,	Para mal, siempre **hay** testigos,
son gente cruel, y fiera	son gente crüel y fiera
los del vando Casteluin.	los del **bando Castelvín**.
Tu lindo gallina en fin.	**Tu** lindo gallina, en fin.
Pluguiera a Dios que estuuiera	¡Pluguiera a Dios que **estuviera**
junto el vando de essa gente,	junto el **bando** de esa gente,
y en aquesta calle armada,	y en aquesta calle armada,
y yo con capa, y espada	y yo con capa y espada
contra todos solamente,	contra todos solamente,
que tu vieras si de alguna	que tú vieras si de alguna
huuiera hazañas tan ciertas;	**hubiera** hazañas tan ciertas,
pero coger entre puertas,	pero coger entre puertas
esso es desgracia perruna.	**eso** es desgracia perruna.
Si tienes tanto desseo	Si tienes tanto **deseo**...
de ver aqueste festin,	de ver aqueste **festín**,
donde el vando Casteluin	donde el **bando Castelvín**
junto, y con cuydado veo,	junto y con **cuidado** veo,
ponte vna mascara, y entra	ponte **una máscara** y entra
pensaran que eres pariente.	**pensarán** que eres pariente.

Figure 1. Original (left) and modernized version by our best model (right) of an excerpt from *Castelvines y Monteses* by Lope de Vega. **Red** words are incorrectly modernized, **green** words are correctly modernized.

Discussion

The motivation for establishing a trained model for the modernisation of historical Spanish came from the expert knowledge of two of the authors of this paper. Both researchers focus on Spanish Golden Age theater and are thus interested in collecting as many Golden Age theatrical pieces as possible with modernized orthography in order to apply stylometric analysis. Beyond that specific motivation, we see the chance that this resource may be used by historians and editors who used to manually transcribe texts produced in the 17th century which were not yet digitized and which are available in cultural heritage institutions, especially libraries and archives. This potential application beyond the narrow focus on Spanish Golden Age theater also marks the (possible) limitations of the current study: while the overall error rate of 4.20% can be regarded as satisfying, the results were evaluated in a first step only on the basis of dramas written in verse form in 17th-Century Spanish.

For testing the model with prose, as a second step, we selected the *Obras completas de Miguel de Cervantes Saavedra* edited by Rodolfo Schevill and Adolfo Bonilla [1914], and converted to electronic text by Fred F. Jehle [1998]. They reproduce the orthography as it appeared in the first edition of the texts, without correcting or modifying it:

> Señalo en las notas las peculiaridades ortográficas sin subsanarlas en el texto, porque el rectificarlas a cada paso parece desnaturalizar la primera edición, dándole un aspecto pulido que desdice enteramente de su carácter. Si se encontrasen estos rasgos en el manuscrito de Cervantes, nadie se atrevería a tocarlos, y, aunque ignoramos con qué fidelidad la primera edición refleja la ortografía del manuscrito, ya que no poseemos éste (vale repetirlo), no es lícito entregarnos a cambios de mero antojo por más limado que resultara el texto. (Schevill and Bonilla, 1914, p. 8)

The overall results of the modernization process seem quite satisfactory, as can be seen in Figure 2 with the first paragraphs of *Don Quixote* (correct modernization marked in green and incorrect in red). Most of the words are correctly modernized. Some errors appear with very infrequent words («*palomino*», «*vellorí*») and context dependent words («*se*» vs «*sé*»). We also calculated the character error rate and the BLEU score of the modernized version by our best model and a modern edition of *Don Quixote* made publicly available by the Biblioteca Virtual Miguel de Cervantes. We compared the metrics obtained by our best model against the scores of a baseline in which no modernization is performed (see Table 2).

Model	BLEU	WER	CER
Baseline (no changes)	50.18	30.95%	10.88%
mT5	61.24	24.56%	11.37%
ByT5	**63.33**	**23.25%**	**10.54%**

Table 2. Scores of the best models on the test set and a baseline with no changes evaluated on *Don Quixote*. Best scores in bold.

En vn lugar de la Mancha, de cuyo nombre no quiero acordarme, no ha mucho tiempo que viuia vn hidalgo de los de lança en astillero, adarga antigua, rozin flaco y galgo corredor. Vna olla de algo mas vaca que carnero, salpicon las mas noches, duelos y quebrantos los sabados, lantejas los viernes, algun palomino de añadidura los domingos, consumian las tres partes de su hazienda. El resto della concluian sayo de velarte, calças de velludo para las fiestas, con sus pantuflos de lo mesmo, y los dias de entre semana se honraua con su vellori de lo mas fino. Tenia en su casa vna ama que passaua de los quarenta, y vna sobrina que no llegaua a los veynte, y vn moço de campo y plaça, que assi ensillaua el rozin como tomaua la podadera. Frisaua la edad de nuestro hidalgo con los cinquenta años. Era de complexion rezia, seco de carnes, enjuto de rostro, gran madrugador y amigo de la caça. Quieren dezir que tenia el sobrenombre de Quixada, o Quesada, que en esto ay alguna diferencia en los autores que deste caso escriuen, aunque por conjeturas verosimiles se dexa entender que se llamaua Quexana. Pero esto importa poco a nuestro cuento; basta que en la narracion del no se salga vn punto de la verdad.	En **un** lugar de la Mancha, de cuyo nombre no quiero acordarme, no ha mucho tiempo que **vivía un** hidalgo de los de **lanza** en astillero, adarga antigua, **rocín** flaco y galgo corredor. **Una** olla de algo más **baca** que carnero, **salpicón** las más noches, duelos y quebrantos los **sabados**, lantejas los viernes, algún **palómino** de añadidura los domingos, **consumían** las tres partes de su **hacienda**. El resto della **concluían** sayo de velarte, **calzas** de velludo para las fiestas, con sus pantuflos de lo mesmo, y los **días** de entre semana se **honraba** con su **bellori** de lo **más** fino. **Tenía** en su casa **una** ama que **pasaba** de los **cuarenta** y **una** sobrina que no **llegaba** a los **veinte** y **un mozo** de campo y **plaza**, que **así ensillaba** el **rocín** como **tomaba** la podadera. **Frisaba** la edad de nuestro hidalgo con los **cincuenta** años. Era de **complejión recia**, seco de carnes, enjuto de rostro, gran madrugador y amigo de la **caza**. Quieren **decir** que **tenía** el sobrenombre de **Quijada** o Quesada, que en esto **hay** alguna diferencia en los autores que deste caso **escriben**, aunque por conjeturas **verosímiles** se deja entender que **sé** **llamaba Quejana**. Pero esto importa poco a nuestro cuento; basta que en la **narración** **de él** no se salga **un** punto de la verdad.

Figure 2. Original (left) and modernized version by our best model (right) of an excerpt from *Don Quixote* by Miguel de Cervantes. **Red** words are incorrectly modernized, **green** words are correctly modernized.

The model can also be useful as a second phase after an OCR-HTR process. In recent years, there has been a growing interest in the creation of models able to transcribe ancient prints and manuscripts using, for instance, resources as Transkribus or eScriptorium (Kahle *et al*. 2017, Mühlberger *et al*. 2019). These models usually maintain the orthography in the document in order to better perform this automatic transcription. It is possible to train transcription models able to automatically modernize the texts (Cuéllar 2021a and 2021b), but this is not common among researchers. It is usually preferred to respect the original orthography and apply a postprocess of modernization. Our model can be helpful in this post-processing step.

However, there is a broad range of orthographic variation (Sebastián Mediavilla 2007), and orthography may differ from one publishing house to another (every publisher may have used their same set of letter sorts), or from one region in which printing presses were located to another (printers' workshops may have exchanged letter sorts amongst themselves within the limits of a specific region). Thus, the modernization of historical texts that were not produced in the same conditions as our corpus may lead to poorer results. Finally, we found slight differences in punctuation and spelling in the dramas with current orthography in the parallel corpus, even though the aim of these editions was to use modern normalized Spanish. This observation opens up the question of how a human reader would evaluate the outputs of our trained model: Would they regard the small "aberrations" from normalized Spanish as idiosyncrasies produced by the transcribers, or would they be able to determine that it was a machine which did the "translation"? Recent research (Clark *et al.* 2021) points out that human readers might not be able to do so, but it was beyond the scope of this study to conduct a comparable assessment. While some of these undesired effects may be addressed by training at the stanza or greater hierarchical level to capture longer range contextual information, it might also imply significantly higher computing resources, training times, and manual revision.

Conclusion

In this work, we have built a parallel corpus of 44 Spanish Golden Age dramas with text in both 17th-Century Spanish and modern Spanish. We have fine-tuned language models on the task of orthographic modernization and show a substantial

improvement of token-free models over token-based models and baseline. We closely analyzed the errors produced and assessed possible causes and mitigation formulas. We are also releasing our best model hoping to foster research within the Spanish Golden Age period and to establish an alternative to the current cumbersome approach of modernizing Golden Age texts solely by hand.

Availability

A demo of our system can be found at <https://huggingface.co/modernisa>.

Our best model can be downloaded from <https://huggingface.co/modernisa/modernisa-byt5-base>.

Acknowledgements

The authors are extremely thankful for the professional input on the legal framework provided by law scholars Dr. Dr. Grischka Petri and Fabian Rack (both FIZ Karlsruhe, Leibniz-Institut für Informationsinfrastruktur). At the same time, the authors would like to recommend their peers to seek advice from the legal helpdesks available at their research institutions or in the country these research institutions are based in.

BIBLIOGRAPHY

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal and Arvind Neelakantan, «Language Models Are Few-Shot Learners», in *Advances in Neural Information Processing Systems*, 33 (2020), pp. 1877-1901.

Cerezo Soler, Juan, and José Calvo Tello, «Autoría y estilo. Una atribución cervantina desde las humanidades digitales. El caso de *La conquista de Jerusalén*», *Anales Cervantinos*, LI (2019), pp. 231-250.

Cervantes Saavedra, Miguel de, *Obras Completas de Miguel de Cervantes Saavedra*, eds. R. Schevill and A. Bonilla y San Martín, Gráficas reunidas, Madrid, 1914.

Clark, Elizabeth, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan and Noah A. Smith, «All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text», in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 7282-7296, online, <https://doi.org/10.18653/v1/2021.acl-long.565>. Accessed on 19th November 2023.

Cuéllar, Álvaro, «Spanish Golden Age Manuscripts (Spelling Modernization) 1.0», *Transkribus*, 2021a.

Cuéllar, Álvaro, «Spanish Golden Age Prints (Spelling Modernization) 1.0», *Transkribus*, 2021b.

Cuéllar, Álvaro, «Cronología y Estilometría: Datación Automática de Comedias de Lope de Vega», *Anuario Lope de Vega. Texto, Literatura, Cultura*, XXIX (2023), pp. 97-130.

Cuéllar, Álvaro, and Germán Vega García-Luengos, «CETSO: Corpus de Estilometría Aplicada al Teatro Del Siglo de Oro», 2017a, online, <http://etso.es/cetso/>. Accessed on 19th November 2023.

Cuéllar, Álvaro, and Germán Vega García-Luengos, «ETSO: Estilometría Aplicada al Teatro Del Siglo de Oro», 2017b, online, <http://etso.es/>. Accessed on 19th November 2023.

Cuéllar, Álvaro, and 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,* XXIX (2023), pp. 131-198.

De la Rosa, Javier, and Juan Luis Suárez, «The Life of *Lazarillo de Tormes* and of His Machine Learning Adversities: Non-Traditional Authorship Attribution Techniques in the Context of the *Lazarillo*», *Lemir: Revista de Literatura Española Medieval y Del Renacimiento*, 20 (2016), pp. 373-438.

Dematтè, Claudia, «Una nueva comedia en colaboración entre ¿Calderón?, Rojas Zorrilla y Montalbán: *Empezar a ser amigos* a la luz del análisis estilométrico», *Rilce. Revista de Filología Hispánica*, XXXV 3 (2019), pp. 852-874.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova, «BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding», in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171-4186, online, <https://doi.org/10.18653/v1/N19-1423>. Accessed on 19th November 2023.

European Parliament and the Council of the European Union, «Database Directive - Council Directive 96/9/EC of 11 March 1996 on the Legal Protection of Databases [1996] OJ L77/20», 1996, online, <https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31996L0009:EN:HTML>. Accessed on 19th November 2023.

Fiore, Arianna, «Questioni di autorialità a proposito di tre commedie seicentesche: Pedro de Urdemalas tra Cervantes, Lope, Montalbán, Diamante e la scuola di Calderón», *Artifara*, XX 2 (2020), pp. 53-77.

Gabay, Simon, Pedro Ortiz Suárez, Alexandre Bartz, Alix Chagué, Rachel Bawden, Philippe Gambette and Benoît Sagot, «From FreEM to D'AlemBERT: A Large Corpus and a Language Model for Early Modern French», 2022, online, <https://doi.org/10.48550/ARXIV.2202.09452>. Accessed on 19th November 2023.

García-Reidy, Alejandro, «Deconstructing the Authorship of Siempre Ayuda La Verdad: A Play by Lope de Vega?», *Neophilologus*, CIII 4 (2019), pp. 493-510.

He, Pengcheng, Xiaodong Liu, Jianfeng Gao and Wei Chen, «DeBERTa: Decoding-Enhanced BERT with Disentangled Attention», in *2021 International Conference on Learning Representations*, 2021, online, <https://www.microsoft.com/en-us/research/publication/deberta-decoding-enhanced-bert-with-disentangled-attention-2/>. Accessed on 19th November 2023.

Jehle, Fred. *Obras Completas de Miguel de Cervantes Saavedra. Texto Electrónico*,

1996, online, <http://www.csdl.tamu.edu/cervantes/>. Accessed on 19th November 2023.

Kahle, Philip, Sebastian Colutto, Günter Hackl and Günter Mühlberger, «Transkribus. A Service Platform for Transcription, Recognition and Retrieval of Historical Documents», in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017, pp. 19-24, online, <https://doi.org/10.1109/ICDAR.2017.307>. Accessed on 19th November 2023.

Liu, Yinhan, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis and Luke Zettlemoyer, «Multilingual Denoising Pre-Training for Neural Machine Translation», *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 726-742, online, <https://doi.org/10.1162/tacl_a_00343>. Accessed on 19th November 2023.

Manjavacas Arévalo, Enrique, and Lauren Fonteyn, «MacBERTh: Development and Evaluation of a Historically Pre-Trained Language Model for English (1450-1950)», in *Proceedings of the Workshop on Natural Language Processing for Digital Humanities (NLP4DH)*, 2021, pp. 23-36, online, <https://aclanthology.org/2021.nlp4dh-1.4.pdf>. Accessed on 19th November 2023.

Manjavacas Arévalo, Enrique, and Lauren Fonteyn, «Non-Parametric Word Sense Disambiguation for Historical Languages», in *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, 2022, pp. 123-134, online, <https://aclanthology.org/2022.nlp4dh-1.16>. Accessed on 19th November 2023.

Manjavacas Arévalo, and Lauren Fonteyn, «Adapting vs. Pre-Training Language Models for Historical Languages», *Journal of Data Mining & Digital Humanities NLP4DH (Digital humanities in...)*, 2022, online, <https://doi.org/10.46298/jdmdh.9152>. Accessed on 19th November 2023.

Mühlberger, Günter, Louise Seaward, Melissa Terras, Sofia Oliveira, Vicente Bosch, Maximilian Bryan and Sebastian Colutto, «Transforming Scholarship in the Archives Through Handwritten Text Recognition: Transkribus as a Case Study», *Journal of Documentation*, LXXV 5 (2019), pp. 954-976.

Papineni, Kishore, Salim Roukos, Todd Ward and Wei-Jing Zhu, «Bleu: A Method for Automatic Evaluation of Machine Translation», in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311-318, online, <https://doi.org/10.3115/1073083.1073135>. Accessed on 19th November 2023.

Post, Matt, «A Call for Clarity in Reporting BLEU Scores», in *Proceedings of the Third Conference on Machine Translation: Research Papers*, Brussels, 2018, pp. 186-191, online, <https://doi.org/10.18653/v1/W18-6319>. Accessed on 19th November 2023.

Real Academia Española and Asociación de Academias de la Lengua Española, eds., *Ortografía de La Lengua Española*, Espasa, Madrid, 2010.

Reynaert, Martin, Maarten van Gompel, Ko van der Sloot and Antal van den Bosch, «PICCL: Philosophical Integrator of Computational and Corpus Libraries: CLARIN Annual Conference 2015», in *Proceedings of CLARIN Annual Conference 2015*, ed. K. De Smedt, 2015, pp. 75-79.

Schweter, Stefan, Luisa März, Katharina Schmid and Erion Çano, «HmBERT: Historical Multilingual Language Models for Named Entity Recognition», in *Proceedings of the Working Notes of CLEF 2022. Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th-to-8th, 2022*, eds. G. Faggioli, N. Ferro, A. Hanbury and M. Potthast, 2022, pp. 1109-1129, online, <http://ceur-ws.org/Vol-3180/paper-87.pdf>. Accessed on 19th November 2023.

Sebastián Mediavilla, Fidel, *Puntuación, humanismo e imprenta en el Siglo de Oro*, Universidad de Vigo-Academia del Hispanismo, Vigo, 2007.

Strubell, Emma, Ananya Ganesh and Andrew McCallum, «Energy and Policy Considerations for Deep Learning in NLP», in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3645-3650, online, <https://doi.org/10.18653/v1/P19-1355>. Accessed on 19th November 2023.

Teatro Español Del Siglo de Oro (TESO), «Editorial Policy», 2022, online, <https://quod.lib.umich.edu/t/teso/ed-policy.html>. Accessed on 19th November 2023.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser and Illia Polosukhin, «Attention Is All You Need», in *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc, 2017, online, <https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>. Accessed on 19th November 2023.

Vega García-Luengos, Germán, «Para la delimitación del repertorio de comedias auténticas de Lope: confirmaciones de autoría y nuevas atribuciones desde La estilometría (II)», *Anuario Lope de Vega. Texto, Literatura. Cultura*, XXIX (2023), pp. 469-544.

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), pp. 91-108.

Vega, Lope de, *Castelvines y Monteses*, Biblioteca Virtual Miguel de Cervantes, Alicante, 2003, online, <https://www.cervantesvirtual.com/nd/ark:/59851/bmc416t9>. Accessed on 19th November 2023.

Xue, Linting, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts and Colin Raffel, «ByT5: Towards a Token-Free Future with Pre-Trained Byte-to-Byte Models», *Transactions of the Association for Computational Linguistics*, X 0 (2022), pp. 291-306.

Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua and Colin Raffel, «MT5: A Massively Multilingual Pre-Trained Text-to-Text Transformer», in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 483-498, online, <https://doi.org/10.18653/v1/2021.naacl-main.41>. Accessed on 19th November 2023.

Appendix A: List of the 44 dramas used as parallel corpus

Pedro Calderón de la Barca
A secreto agravio, secreta venganza
Afectos de odio y amor
Amado y aborrecido
Amor, honor y poder
Antes que todo es mi dama
Casa con dos puertas mala es de guardar
Céfalo y Pocris
Celos aun del aire matan
Eco y Narciso
El alcalde de Zalamea
El Faetonte
El fiera, el rayo y la piedra
El galán fantasma
El jardín de Falerina
El mágico prodigioso
El médico de su honra
El monstruo de los jardines
El pintor de su deshonra
El sitio de Bredá
Fieras afemina amor
La aurora en Copacabana
Las cadenas del Demonio
La dama duende
La devoción de la cruz
La vida es sueño
Las fortunas de Andrómeda y Perseo
Mañanas de abril y mayo
No hay burlas con el amor

Félix Lope de Vega Carpio
Al pasar del arroyo
Amar, servir y esperar
El amigo por fuerza
El conde Fernán González
Fuenteovejuna
Los amantes sin amor

Juan Pérez de Montalbán
Don Florisel de Niquea
El mariscal de Virón
La doncella de labor
Los amantes de Teruel
Olimpia y Vireno

Juan Ruiz de Alarcón
El examen de maridos
El tejedor de Segovia
Las paredes oyen
Los pechos privilegiados
Ganar amigos