

ArDraCor: a new corpus of Argentine nineteenth-century dramatic texts

Ulrike Henny-Krahmer¹ ([ORCID](#)), Gimena del Rio Riande² ([ORCID](#)), Laura Volkind² ([ORCID](#)), Romina de León² ([ORCID](#)), Nidia Hernández² ([ORCID](#)), María Teresa Ravelo Sánchez³ ([ORCID](#)), Erik Renz¹ ([ORCID](#))

1: Universität Rostock, Germany

2: Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina

3: Universidad Nacional Autónoma de México (UNAM), Mexico

With our contribution, we aim to present a new initiative for creating a corpus of Argentine nineteenth-century dramatic texts called ArDraCor (*Argentine Drama Corpus*). This will be the first DraCor corpus with Latin American texts and will therefore contribute to the diversity of corpora on the platform. Culturally and spatially, the corpus covers drama written by Argentine authors and dramatic texts that were first published in Argentina during the nineteenth century. Temporally, we consider the period of the long nineteenth century and include works that were premiered or first published between 1780 and 1920. Initially, the corpus is being created by a team of researchers from CONICET in Argentina, the University of Rostock in Germany and the UNAM in Mexico. The texts are encoded in XML-TEI and the corpus will be published at <https://github.com/dracor-org/ardracor>.

We are currently in the initial phase of compiling the corpus. So far, almost 100 drama texts have been identified that are suitable for the corpus and are already available in the form of digitized images and, in some cases, in HTML. The main sources for the texts are the Biblioteca Virtual Miguel de Cervantes, the collection of texts of the Academia Argentina de Letras on Wikimedia Commons and resources that are available via the Instituto Nacional del Teatro. There is not yet a single text in XML-TEI format, so we have to set up a workflow that includes text recognition using OCR and converting existing OCRed text or other formats into TEI. Currently, we are using the OCR4all tool (Reul et al. 2019, see also Dennerlein et al. 2025) and an encoding in the EzDrama format (Skorinkin 2024) to prepare the digital full texts for encoding in TEI.

By the time of the DraCor Summit, we will have encoded the first texts in TEI and published them in the repository on GitHub. Besides presenting the first results of our new corpus ArDraCor, we hope that the summit will provide us with feedback on our corpus design and workflow, which we can then take into account in the further development.

References

Dennerlein, Katrin, Martin Rupnig, and Christian Reul. 2025. "Zum Aufbau digitaler Dramenkorpora. OCR4alltoDraCorTEI als Baustein für die Edition von maschinenlesbaren Versionen historischer Dramendrucke." In *DHd2025. Under Construction. Book of Abstracts*. Zenodo. <https://doi.org/10.5281/zenodo.14942992>.

Reul, Christian, Dennis Christ, Alexander Hartelt, Nico Balbach, Maximilian Wehner, Uwe Springmann, Christoph Wick, Christine Grundig, Andreas Büttner, and Frank Puppe. 2019. "OCR4all — An open-source tool providing a (semi-) automatic OCR workflow for historical printings." *Applied Sciences* 9 (22). <https://doi.org/10.3390/app9224853>.

Skorinkin, Daniil. 2024. "EasyDrama: a lightweight solution for encoding plays in TEI/XML." In: *TEI 2024 Book of Abstracts*, edited by Gimena del Rio Riande, 38–39. Buenos Aires: Universidad del Salvador. <https://tei2024.tei-c.org/pdf/tei2024boa.pdf>. Accessed May 6, 2025.