

Distributed Corpus Building in Literary Studies: The DraCor Example

Giovannini, Luca

giovannini@uni-potsdam.de
Universität Potsdam

Skorinkin, Daniil

skorinkin@uni-potsdam.de
Universität Potsdam

Trilcke, Peer

trilcke@uni-potsdam.de
Universität Potsdam

Börner, Ingo

ingo.boerner@uni-potsdam.de
Universität Potsdam

Fischer, Frank

fr.fischer@fu-berlin.de
Freie Universität Berlin

Dudar, Julia

dudar@uni-trier.de
Universität Trier

Milling, Carsten

milling@uni-potsdam.de
Universität Potsdam

Pořízka, Petr

petr.porizka@upol.cz
Palacký University Olomouc

Introducing DraCor

The DraCor project, based on the concept of “Programmable Corpora” (Fischer et al. 2019), is an open platform as well as a growing network of resources for hosting, accessing, and analysing theatre plays. Presently including 15 corpora in 10 different languages, totalling about 3000 works, it provides scholars not only with a wealth of TEI-encoded digital texts, but also with applications and tools for various research purposes – ranging from the computation of textual network metrics via extraction functions to SPARQL queries and speech distribution statistics.

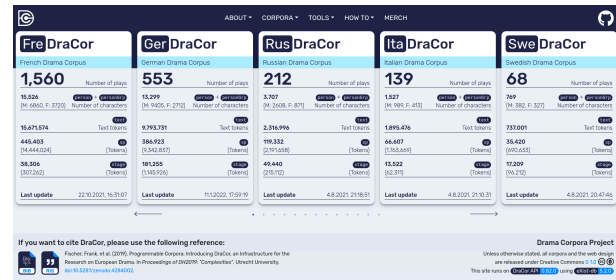


Figure 1. DraCor homepage (www.dracor.org)

Since its corpora are created either by aggregating formally heterogeneous texts from different sources or by transforming existing collections, there’s no standard path for the ingestion of dramatic texts into DraCor. An overview of the main pipelines is presented in Figure 2. While the current workflow poses some technical challenges in terms of corpus homogenisation, which are currently being tackled through the prototyping of several tools, it also highlights the collaborative, community-driven nature of our endeavours.

Indeed, by relying on the general TEI-P5 model for dramatic texts, with minimal enhancements, DraCor strives to facilitate contributions by external scholars who want to onboard their corpora onto its ecosystem. Adopting a distributed and decentralised corpus building approach, we aim at building a community of practice around the project, bringing together domain (e.g. literary scholars) and technical experts (developers) to contribute to the growth of our collections.

To achieve this aim, we are currently developing tools and resources to make the corpus-building process as transparent and intuitive as possible, while establishing some standards in terms of ontologies, metadata treatment, editorial conventions and scripts. On one side, we are working on formalising our know-how by producing some guidelines and tutorials.¹ On the other side, we are currently developing a lightweight Markdown language for semi-automatic conversion of raw texts into basic DraCor XMLs (*Ez-Drama*),² catering to the needs of scholars without technical background and empowering them to easily prepare new texts for onboarding.

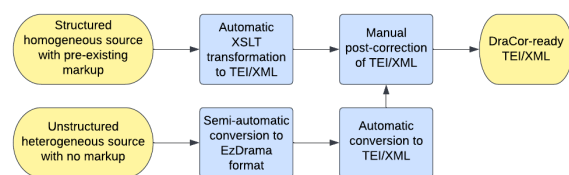


Figure 2. DraCor onboarding pipeline.

Examples of Collaborative Corpus Building

From its inception, the DraCor project has been open to small and large corpora of plays from any linguistic and cultural background, encouraging external scholars to prepare and submit their own texts. While the first, prototypical corpora, such as the Ger-

man and the Russian ones, were built by the DraCor team, soon new collections were born out of community-driven efforts. Early examples included the Spanish, Hungarian, and Alsatian corpora – with the latter being notable for its low-resources approach (Ruiz Fabo et al. 2022). Thanks to the guidelines and the scripts previously described, the pipeline for onboarding has now been refined, allowing for efficient distributed work on multiple sets of texts. As a showcase, we present here some corpora currently in production.

Originating from a collection of heterogeneous online sources without previous markup, the Ukrainian Drama Corpus (UDraCor) has represented a test ground for the development and deployment of the *EzDrama* conversion script. At the same time, UDraCor has shown again the strengths of a community-based corpus-building approach, insofar as it has involved scholars specialising in Ukrainian studies and volunteers in the text collection and encoding phases.

A similar approach has been followed in the development of the Czech Drama Corpus (CzeDraCor), started with the pilot encoding of 10 plays by Karel and Josef Čapek. In this case, funding from CLS INFRA has allowed a research fellow to work full-time on the project. After collecting the plays' texts and checking them against standard scholarly Čapek editions, they will be processed through the *EzDrama* parser and manually revised before integration into DraCor.

While maintaining a non-Anglocentric approach, the DraCor project is also looking forward to extending its English-language collections. Accordingly, we are processing about 800 plays from the Early Print collection (EPDraCor) through a semi-automated workflow including XSLT transformations, metadata enrichment through OpenRefine and rounds of manual correction to solve character identification issues (using our prototypical “Who-Is-@who” disambiguation tool ³).

While these three corpora are still being assembled, plans for new additions to the DraCor library (in Polish and Hebrew) are already being discussed, and further proposals are welcome. Accordingly, we hope to use this DH2023 conference as a platform for showcasing the open nature of DraCor and inviting colleagues from all backgrounds to consider contributing to it.

Acknowledgments

In the context of CLS INFRA (<https://clsinfra.io/>), DraCor has received funding from the European Union's Horizon 2020 program (grant agreement No. 101004984). Bohdan Tokarskyi (University of Potsdam) helped to establish an initial list of 152 Ukrainian plays (1813–1948) for UDraCor.

Notes

1. <https://dracor.org/doc/tutorials>
2. <https://github.com/dracor-org/ezdrama>
3. See the repository (<https://github.com/dracor-org/epdra-cor-whois>) and the interface (<https://dracor-org.github.io/epdra-cor-whois>).

Bibliography

Fischer, Frank / Ingo Börner / Mathias Göbel / Angelika Hecht / Christopher Kittel / Carsten Milling / Peer Trilcke

(2019): Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. *Proceedings of DH2019: Complexities*, Utrecht, The Netherlands, July 2019. <https://doi.org/10.5281/zenodo.4284002> [29.04.2023].

Ruiz Fabo, Pablo / Delphine Bernhard / Andrew Briand / Carole Werner (2022): Computational drama analysis from almost zero electronic text: The case of Alsatian theater. *Computational Drama Analysis: Achievements and Opportunities*, Cologne, Germany, September 2022. <https://hal.archives-ouvertes.fr/hal-03762377> [29.04.2023].