Onboard onto DraCor.
Prototyping Workflows to Homogenize Drama Corpora for an Open Infrastructure

Börner, Ingo
ingo.boerner@uni-potsdam.de
Universität Potsdam

Fischer, Frank
fr.fischer@fu-berlin.de
Freie Universität Berlin

Giovannini, Luca
giovannini@uni-potsdam.de
Universität Potsdam

Lu, Christopher
christopher.lu@balliol.ox.ac.uk
University of Oxford

Milling, Carsten
milling@uni-potsdam.de
Universität Potsdam

Skorinkin, Daniil
daniil.skorinkin@uni-potsdam.de
Universität Potsdam

Sluyter-Gäthje, Henny
henny.sluyter-gaethje@uni-potsdam.de
Universität Potsdam

Trilcke, Peer
trilcke@uni-potsdam.de
Universität Potsdam

Approaches to Corpus Homogenization

Comparative endeavors in Computational Literary Studies typically require corpora which are both diverse, i.e., including texts in different languages and from different sources, and homogenized, i.e., formal and structural consistent. One way to tackle this issue is to establish upstream internal guidelines, such as the ones developed within the ELTeC initiative (Schöch et al. 2021). [1] In the following, we report on our approach to homogenizing corpora for DraCor. [2]

DraCor, based on the concept of Programmable Corpora (Fischer et al. 2019), is an open platform as well as a growing network for hosting, accessing, and analyzing theater plays. DraCor relies on the general TEI model for dramatic texts, with minimal enhancements, and thus facilitates contributions by external scholars who want to onboard their corpora onto its ecosystem. Once integrated, corpora can benefit from the platform's APIs and services, ranging from the computation of network metrics via various extraction functions to SPARQL queries.

Typically, corpora for DraCor are not built from scratch, but are created either by aggregating formally heterogeneous texts from different sources or by transforming existing corpora. Unlike in ELTeC, the homogenization of texts for DraCor usually does not stand at the beginning of the corpus creation process, but is rather an intervention in existing corpora which are sometimes subject to amendment and growth, hence ›living‹. This approach poses a number of challenges, for which we are currently prototyping several workflows. Here, we present the pipelines for mounting to DraCor two new corpora: the English-language *EarlyPrint Drama Corpus* (EPDraCor) and the *Ukrainian Drama Corpus* (UDraCor). [3]

Corpus Onboarding

From a technical point of view, onboarding corpora onto DraCor is a series of automated and manual transformations of the source data, which depend crucially on the format and markup of the files. Texts from a single, homogeneous collection with pre-existing markup and metadata will require different workflows and pipelines than those coming, for example, from a variety of raw text sources.

This heterogeneous point of departure is what shapes our onboarding approach. Consequently, we are developing a modular workflow made up of a set of demand-dependent components. In addition to guideline-based manual revisions (e.g. pre-structuring texts with Markdown), we use XSLT scripts for automated transformations. Edits specific to theater plays, such as the task of speaker identification, are supported by an Oxygen framework; [4] we are furthermore experimenting with task-specific GUI applications based on react.js. [5] The correction and enrichment of metadata, such as the addition of Wikidata ID, is organized semi-automatically via OpenRefine.

A particular challenge is posed by living corpora. Here, the manual transformations performed during onboarding should be reapplicable in case of edits to the source data. Accordingly, we implemented routines for a ›backward compatibility‹ of the markup: the changes made by us during onboarding can later be applied again to a newer version of the source files. [6]

EPDraCor and UDraCor

To develop our workflows and pipelines, two corpora with very different requirements are currently in the process of onboarding. While UDraCor originates from a gro-

wing collection of heterogeneous sources, EPDraCor is based on semantically rich TEI files from the Early Print project. [7] The onboarding of EPDraCor starts with enhancing and correcting the original markup in our copy of the source corpus, accompanied by collecting LOD metadata from additional sources. Then, we combine the enhanced sources with their metadata and use XSLT to transform them, so that the TEI fulfills the requirements of the DraCor platform.

Due to the heterogeneity of the sources, a more case-specific solution must be found for UDraCor in this initial step. Here, the conversion is a semi-automatic procedure with heavy use of string patterns and regex. At the same time, UDraCor takes a community-based corpus-building approach by inviting scholars specializing in Ukrainian studies to work on both technical and content-related tasks. This work on UDraCor once again shows how the technical task of corpus building and community activities are crucially intertwined.

Fußnoten

1. See https://distantreading.github.io/ELTeC and https://distantreading.github.io/Schema/eltec-1.html.
2. https://dracor.org. 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).
3. Both corpora are still a work in progress. For review, the two corpora can (as public alpha) be accessed in the corresponding GitHub repositories https://github.com/dracor-org/epdracor and https://github.com/dracor-org/udracor. Both corpora will be published as public beta in the context of DHd2023.
4. https://github.com/dracor-org/dracor-oxygen-framework.
5. See e.g. our prototype of a Who-Is-Identification-Tool https://github.com/dracor-org/epdracor-whois and its interface https://dracor-org.github.io/epdracor-whois.
6. For this, see our prototype script in the EPDraCor repository: https://github.com/dracor-org/epdracor.
7. https://earlyprint.org.

Bibliographie

Fischer, Frank, Ingo Börner, Mathias Göbel, Angelika Hechtl, Christopher Kittel, Carsten Milling, and Peer Trilcke. 2019. “Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama”. In *Proceedings of DH2019: "Complexities"*. Utrecht: Utrecht University. https://doi.org/10.5281/zenodo.4284002.

Mueller, Martin, and Joseph Loewenstein (eds.). n.d. “Early Print Library”. Accessed August 3, 2022. https://earlyprint.org.

Schöch, Christof, Roxana Patraș, Diana Santos, and Tomaž Erjavec. 2021. "Creating the European Li-terary Text Collection (ELTeC): Challenges and Per-spectives". *Modern Languages Open* 1: 25. https://doi.or-g/10.3828/mlo.v0i0.364.