

Luca Giovannini
(Potsdam/Padua)

Miriam Schirmer
(TU Munich/Regensburg)

The Courtroom as Stage: Exploring the Performative Dimension of Genocide Trials

Cultural Data Analytics Conference
Tallinn, 14.12.2023

Research idea

- Key concept: “the trial is preeminently a theatrical form” (Susan Sontag)
- Some previous research on ‘trials as drama’: Nuremberg and Jerusalem Nazi trials, South African apartheid commission
- Mostly non-quantitative and lacking a computational approach!

Research goal

We want to model and analyse transcripts from genocide trials like they were dramatic plays, using infrastructure and tools we borrow from the *Drama Corpora* (*DraCor*) project.

Index

1. Introducing the DraCor ecosystem
2. Building the Genocide Trials Transcripts Corpus (GenoCor)
3. Outlining possible research applications

This presentation: **plu.sh/genocor**

Project repository: **github.com/lucagiovannini7/genocor**

1. Introducing the DraCor ecosystem

What's DraCor?

an open-access collection of
TEI-encoded theatre plays
(15 programmable corpora,
12 languages, +3000 texts)



a research infrastructure for
computational literary studies

a series of tools for
quantitative analysis
(API - SPARQL endpoints -
Python + R packages -
Docker)

a community of researchers
and developers



Fre DraCor

French Drama Corpus

1,560

Number of plays

15,526

person + **personGrp**

(M: 6860, F: 3720) Number of characters

text

15,671,574

Text tokens

445,403

sp

(14,444,024)

(Tokens)

38,306

stage

(307,262)

(Tokens)

Last update

22.10.2021, 16:31:07

Ger DraCor

German Drama Corpus

553

Number of plays

13,299

person + **personGrp**

(M: 9405, F: 2712) Number of characters

text

9,793,731

Text tokens

386,923

sp

(9,342,837)

(Tokens)

181,255

stage

(1,145,926)

(Tokens)

Last update

11.1.2022, 17:59:19

Rus DraCor

Russian Drama Corpus

212

Number of plays

3,707

person + **personGrp**

(M: 2608, F: 871) Number of characters

text

2,316,996

Text tokens

119,332

sp

(2,191,658)

(Tokens)

49,440

stage

(215,112)

(Tokens)

Last update

4.8.2021, 21:18:51

Ita DraCor

Italian Drama Corpus

139

Number of plays

1,527

person + **personGrp**

(M: 989, F: 413) Number of characters

text

1,895,476

Text tokens

66,607

sp

(1,763,669)

(Tokens)

13,522

stage

(62,311)

(Tokens)

Last update

4.8.2021, 21:10:31

Swe DraCor

Swedish Drama Corpus

68

Number of plays

769

person + **personGrp**

(M: 382, F: 327) Number of characters

text

737,001

Text tokens

35,420

sp

(690,633)

(Tokens)

17,209

stage

(96,212)

(Tokens)

Last update

4.8.2021, 20:47:46



If you want to cite DraCor, please use the following reference:



Fischer, Frank, et al. (2019). Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. In *Proceedings of DH2019: "Complexities"*, Utrecht University, doi:10.5281/zenodo.4284002.

Drama Corpora Project

Unless otherwise stated, all corpora and the web design are released under Creative Commons [0 1.0](#)

This site runs on [DraCor API 0.82.0](#) using [eXist-db 5.2.0](#)

DraCor landing page: <https://dracor.org>

Corpora currently in DraCor





German Drama Corpus

Edited by Frank Fischer and Peer Trilcke. Features more than 500 German-language plays from the 1730s to the 1940s. For a corpus description and full credits please see the [README on GitHub](#).

Corpus licensed under [CC0](#)

Download a comprehensive table with metadata on all plays in the corpus: [JSON](#) [CSV](#)

Authors ▲▼	Title ▲▼	Year (normalized) ▲▼ ⓘ	Network Size ▲▼	Source ▲▼	ID ▲▼
Gryphius, Andreas Q77214	Leo Armenius oder Fürsten-Mord Trauer-spiel Q1818468	1650 1646 1651 1650	37	TextGrid Repository TEI version	ger000546
Gryphius, Andreas Q77214	Horribilicribrifax Teutsch Scherzspiel Q20984917	1650 1647-1650 1674 1663	21	TextGrid Repository TEI version	ger000549
Gryphius, Andreas Q77214	Carolus Stuardus Trauerspiel Q42187688	1650 1649 1650 1657	55	TextGrid Repository TEI version	ger000550
Gryphius, Andreas Q77214	Catharina von Georgien oder Bewehrte Beständigkeit Trauerspiel Q1736032	1651 1647 1651 1657	22	TextGrid Repository TEI version	ger000548

Landing page of GerDraCor: <https://dracor.org/ger>

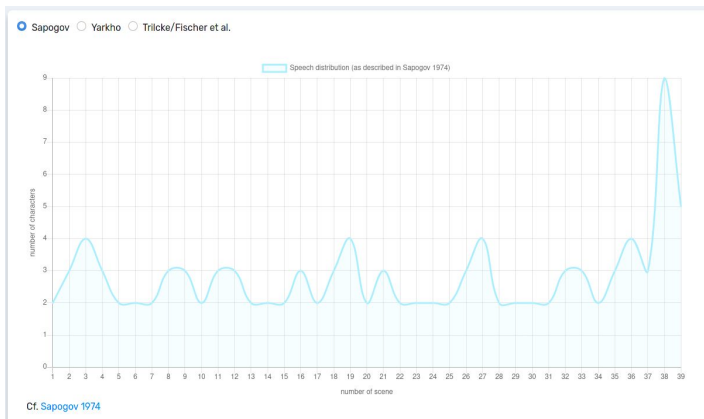
```

-<?xml-model
  href="https://dracor.org/schema.rng" type="application/xml" schematypens="http://relaxng.org/ns/structure/1.0"
?>
-<TEI xml:id="ger000343" xml:lang="ger">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title type="main">Professor Bernhardi</title>
        <title type="sub">Komödie in fünf Akten</title>
      <author>
        <persName>
          <forename>Arthur</forename>
          <surname>Schnitzler</surname>
        </persName>
        <idno type="wikidata">Q44331</idno>
        <idno type="pnd">118609807</idno>
      </author>
    </titleStmt>
    <publicationStmt>
      <publisher xml:id="dracor">DraCor</publisher>
      <idno type="URL">https://dracor.org</idno>
    <availability>
      <licence>
        <ab>CC0 1.0</ab>
        <ref target="https://creativecommons.org/publicdomain/zero/1.0/">Licence</ref>
      </licence>
    </availability>
  </publicationStmt>
  <sourceDesc>
    <bibl type="digitalSource">

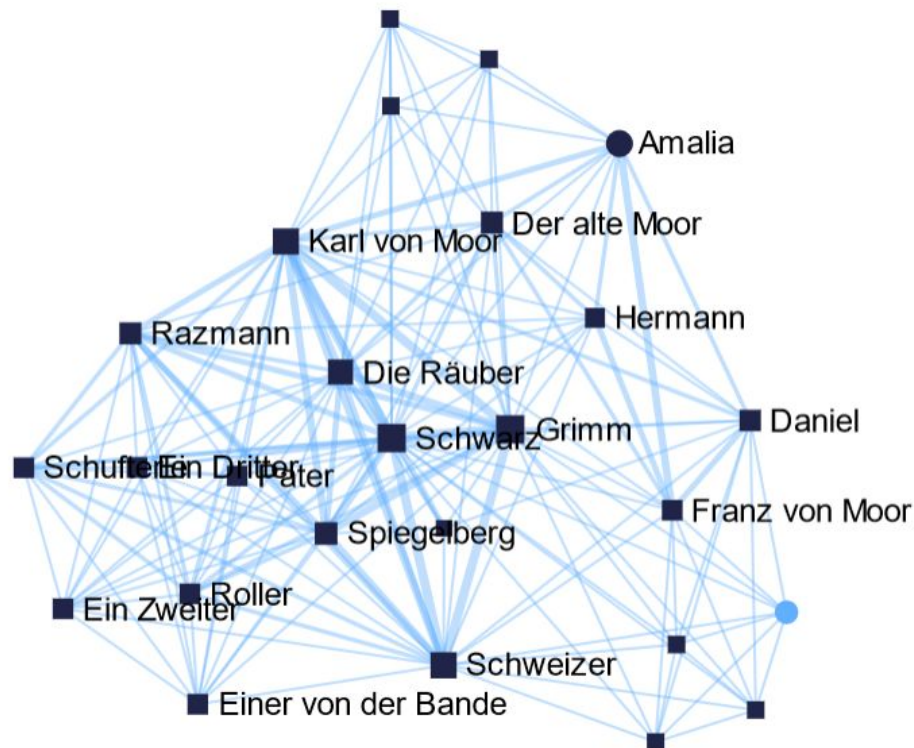
```

Under the hood: XML-TEI files

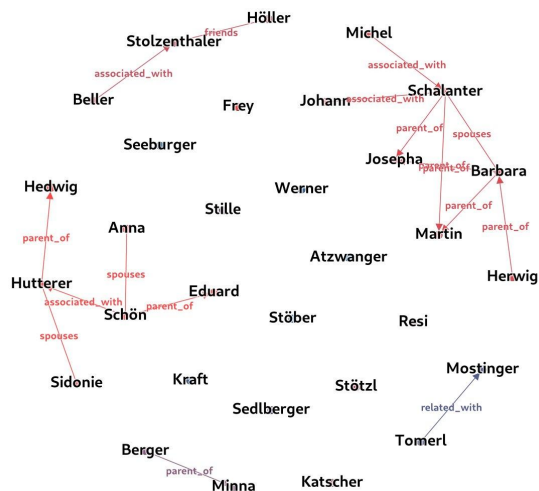
Example: Schnitzler's [*Professor Bernhardi*](#)

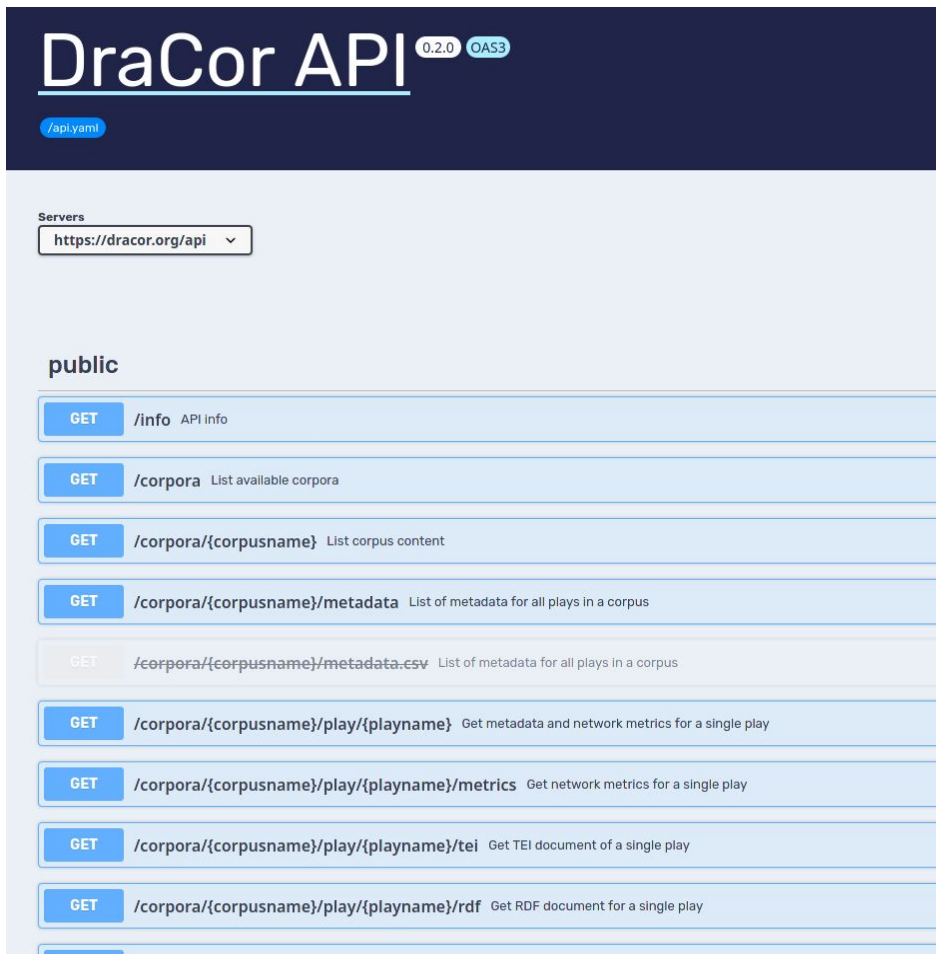


Visualising data in DraCor



Clockwise, from bottom left: L. Gottsched's [The Last Will](#),
 Anzgruber's [The Fourth Commandment](#), Schiller's [The Robbers](#)





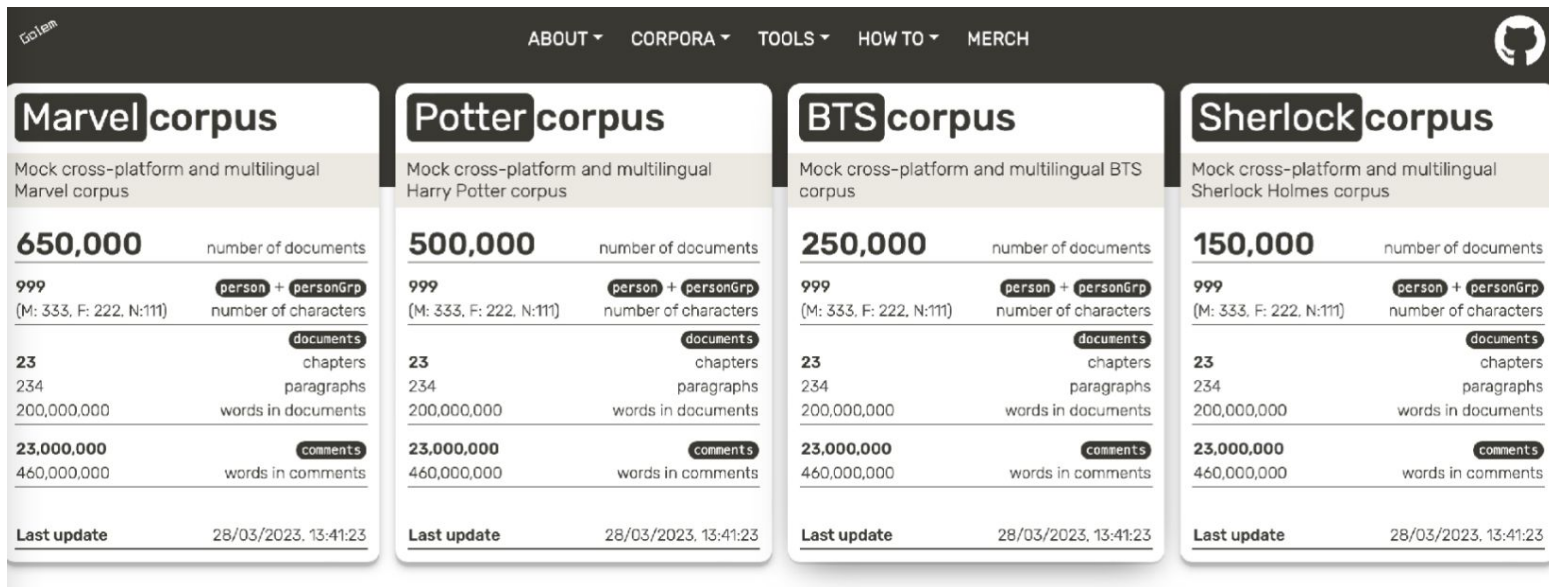
DraCor API

- first stable version (1.0.0) released this month!
- doing direct queries (e.g. asking for speech text or network data)
- getting the corpora metadata table

*... could we build
a DraCor-based corpus
consisting of transcripts
from genocide trials?*

Previous DraCor spin-offs

- In computational literary studies: PoeCor and EcoCor prototypes
- In cultural analytics: GOLEM project (Pianzola et al. 2023)



The screenshot shows the Golem project website with a dark header containing navigation links: ABOUT, CORPORA, TOOLS, HOW TO, and MERCH. A GitHub logo is in the top right corner. Below the header are four white cards, each representing a different corpus. Each card has a title, a description, and a table of statistics. The statistics table for each corpus includes: number of documents, number of characters (with a breakdown by gender/age), number of chapters, paragraphs, words in documents, words in comments, and the last update date.

Corpus	Number of documents	Number of characters (M: 333, F: 222, N:111)	Chapters	Paragraphs	Words in documents	Words in comments	Last update
Marvel corpus	650,000	999	23	234	200,000,000	23,000,000	28/03/2023, 13:41:23
Potter corpus	500,000	999	23	234	200,000,000	23,000,000	28/03/2023, 13:41:23
BTS corpus	250,000	999	23	234	200,000,000	23,000,000	28/03/2023, 13:41:23
Sherlock corpus	150,000	999	23	234	200,000,000	23,000,000	28/03/2023, 13:41:23

2. Building the GenoCor corpus

Case study: Rwanda and Yugoslavia genocides

ICTR (International Criminal Tribunal for Rwanda)

Purpose: Established to prosecute individuals responsible for genocide and other serious violations of international humanitarian law during the 1994 Rwandan genocide.

Background: The Rwandan genocide involved the systematic killing of approximately 800,000 ethnic Tutsis by ethnic Hutus within a span of 100 days.

ICTY (International Criminal Tribunal for the former Yugoslavia)

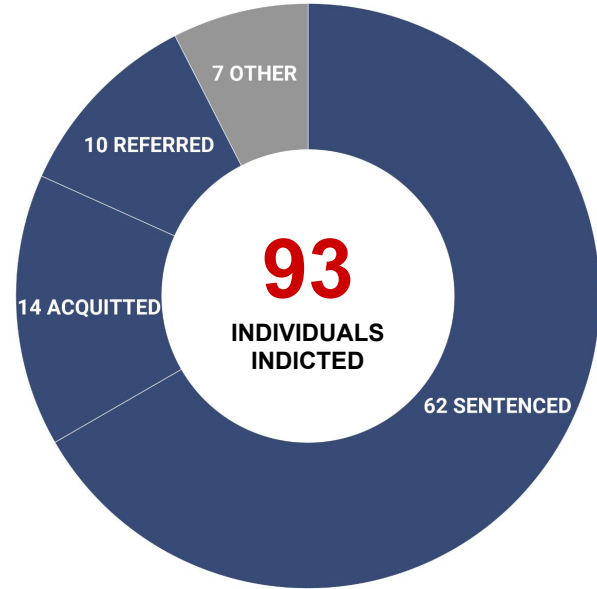
Purpose: Formed to address war crimes that occurred during the conflicts in the Balkans from 1992 to 2001.

Background: The Yugoslav Wars involved ethnic and territorial disputes, resulting in widespread atrocities, ethnic cleansing, and genocide, notably in Bosnia and Herzegovina (Srebrenica).

Factsheet: Indictments

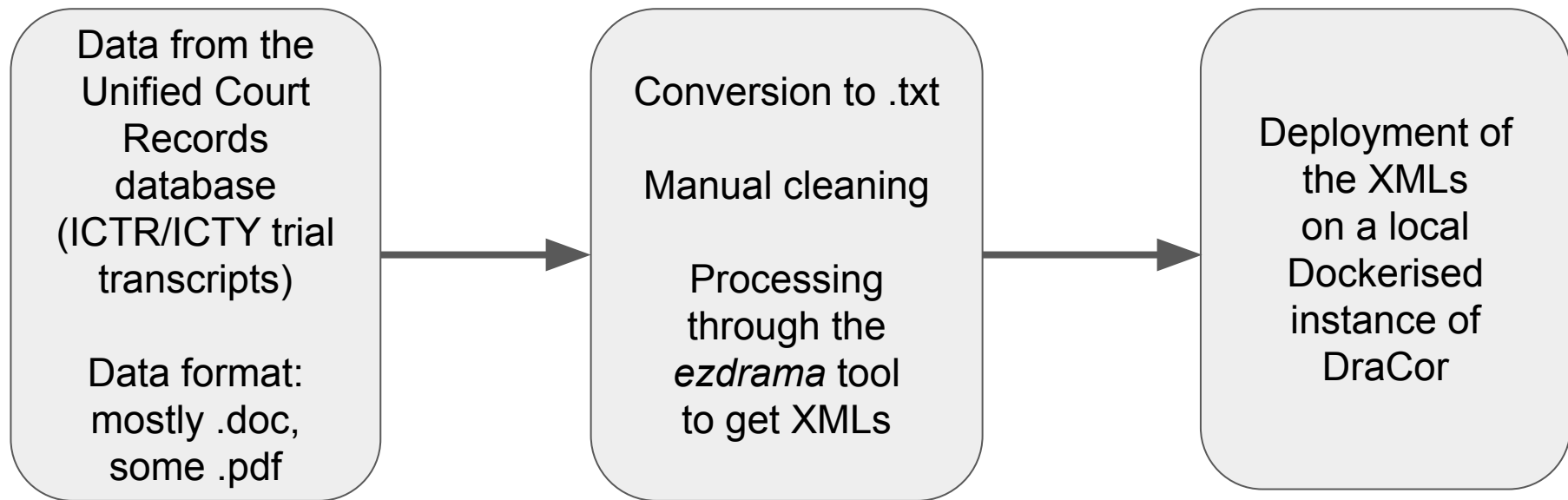


ICTY ([source](#))



ICTR ([source](#))

Corpus building pipeline



Data source



SELECTED DOCUMENTS

English

NZABONIMANA, CALLIXTE (ICTR-98-44D)

NZABONIMANA, CALLIXTE (ICTR-98-44D)

Trial Chamber

Appeals Chamber

President

Others

Indictment
Transcripts

Decisions
Orders
Judgement

Decisions
Orders
Judgement

President's Decisions
President's Orders

Registry Decisions
Submissions and
Requests
Pre-Trial Briefs
Final Briefs
Motions
Warrant Arrest

filter documents

Case interface

Transcripts overview

Transcripts

2011

April

05 04

March

28 23

February

15 02

January

26 25 24 21 20 19 17 11 10

2010

November

25 25 24 22 16 11 10 09 08

October

28 27 26 25 21 20 19 19 18
18 13 12

September

30 29 28 27 23 22 21 20 16
15 14 08 07 06 02

August

30 26 25 24

July

14 13 12 07 06 05

June

30 29 28 25 24 08 07 03
02 01

May

31 27 27 26 26 25 25 24 24
20 20 20 19 19 18 17 17 13

April

29 28 27 26 26 23 22 22 21
21 20 19 19 15 15 14 14 14

1 Tuesday, 31 October 2006

2 [Open session]

3 [The accused entered court]

4 --- Upon commencing at 9.13 a.m.

5 JUDGE AGIUS: Madam Registrar, good morning to you. Could you
6 kindly call the case, please.

7 THE REGISTRAR: Good morning, Your Honours. This is case number
8 IT-05-88-T, the Prosecutor versus Vujadin Popovic et al.

9 JUDGE AGIUS: I thank you, Madam.

10 All accused are here. Defence teams are all here, minus Mr.
11 Bourgon, for justified reason communicated to Chambers.

12 Prosecution, Mr. Vanderpuye and Mr. McCloskey.

13 So I think we can safely start.

14 First thing before I ask if there are any preliminaries or

Sample
transcript
(.doc)

Conversion workflow via the *ezdrama* tool

[Open session]
[The accused entered court]
JUDGE AGIUS: Madam Registrar, good morning to you. Could you kindly call the case, please.
THE REGISTRAR: Good morning, Your Honours. This is case number IT-05-88-T, the Prosecutor versus Vujadin Popovic et al.

\$[Open session]
\$[The accused entered court]
@JUDGE AGIUS: \n
Madam Registrar, good morning to you. Could you kindly call the case, please.
@THE REGISTRAR: \n
Good morning, Your Honours. This is case number IT-05-88-T, the Prosecutor versus Vujadin Popovic et al.

```
<div>
  <stage>
    <p>[Open session]</p>
    <p>[The accused entered court]</p>
  </stage>
  <sp who="#judge-agius">
    <speaker>JUDGE
AGIUS</speaker>
    <p>Madam Registrar, good morning
to you. Could you kindly call the case,
please.</p>
  </sp>
  <sp who="#registrar">
    <speaker>THE
REGISTRAR</speaker>
    <p>Good morning, Your Honours.
This is case number IT-05-88-T, the
Prosecutor versus Vujadin Popovic et
al.</p>
  </sp>
</div>
```

original plain text

application of *ezdrama*
markup via regexes

DraCor-ready XML
produced by *ezdrama* script

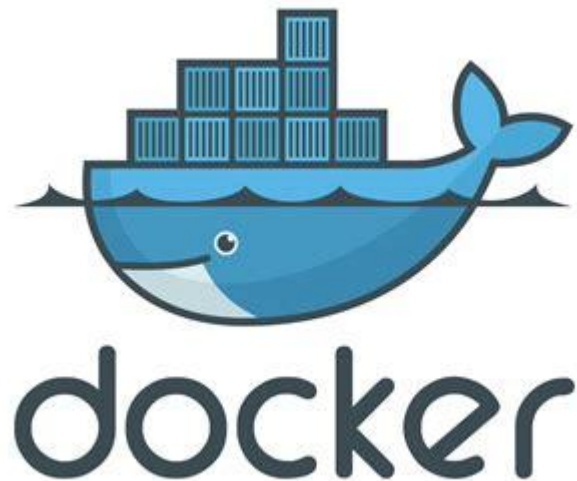
*Trial transcripts **are not** plays, but they might be similar enough to be modelled according to the **same TEI guidelines** and processed through the **same (DraCor) algorithms***

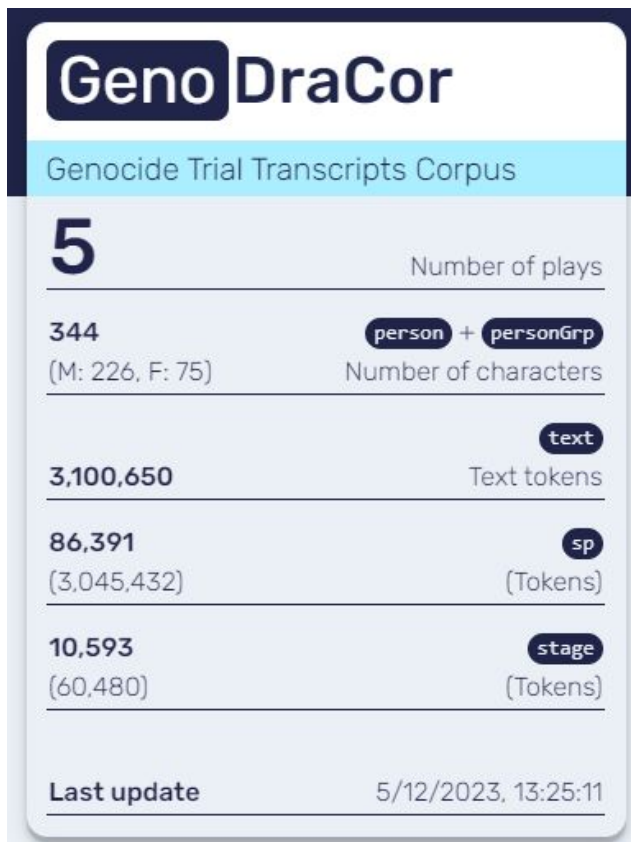
Modelling transcripts as drama: issues

- paratextual materials:
 - list of participants → dramatis personae (cast list)
 - sidenotes → stage directions
- normalisation and disambiguation of speakers' names:
 - (e.g. "Q", "A", "THE WITNESS" → actual names of the speakers)
- segmentation:
 - acts = individual tribunal session (i.e., one transcript)
 - scenes = debate phases (preliminary matters, examination-in-chief, cross-examination, re-examination, pre-trial/status conferences, judgement, appeal hearing)

Dockerising GenoCor

1. We use preloaded Docker images to build a copy of the **whole** DraCor infrastructure on our local machine
 2. We load the GenoCor XMLS into this Dockerised version
- More info: Börner et al. 2023 (see also the [Python package under development](#))





Introducing GenoCor v0.9

(ICTR-only)



Genocide Trial Transcripts Corpus

This is a corpus of transcripts from the Rwandan and Yugoslavian genocide trials (ICTR/ICTY).

Download a comprehensive table with metadata on all plays in the corpus: [JSON](#) [CSV](#)



Authors ▲▼	Title ▲▼	Year (normalized) ▲▼ ⓘ	Network Size ▲▼	Source ▲▼	ID ▲▼
International Criminal Tribunal for Rwanda  Q478358	ICTR-95-1B-I v. Mika Muhimana	1999  1999	90	Court records TEI version	geno000002
International Criminal Tribunal for Rwanda  Q478358	ICTR 2001-71 v. Emmanuel Nindabahizi	2001  2001	41	Court records TEI version	geno000003
International Criminal Tribunal for Rwanda  Q478358	ICTR-2001-64-I v. Sylvestre Gacumbitsi	2001  2001	63	Court records TEI version	geno000001
International Criminal Tribunal for Rwanda  Q478358	ICTR-2001-66-I v. Athanase Seromba	2001  2001	69	Court records TEI version	geno000005
International Criminal Tribunal for Rwanda  Q478358	ICTR-97-31-I v. Tharcisse Renzaho	2005  2005	67	Court records TEI version	geno000004

If you want to cite DraCor, please use the following reference:

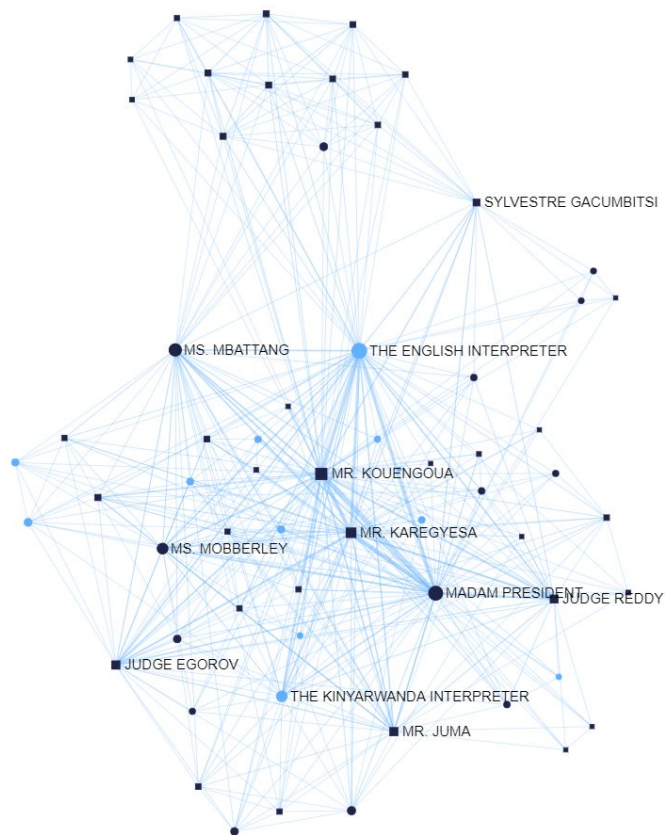


Fischer, Frank, et al. (2019). Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. In *Proceedings of DH2019: "Complexities"*, Utrecht University, [doi:10.5281/zenodo.4284002](https://doi.org/10.5281/zenodo.4284002).

Drama Corpus Project

Unless otherwise stated, all corpora and the web design are released under Creative Commons [0 1.0](#)  

This site runs on [DraCor API 0.90.1-2-g19a3f46-dirty](#) using [eXist-db 6.0.1](#)



This tab shows a co-occurrence network. If characters appear in the same scene or act, they are linked.

Network properties

Segments: 58
All-in at segment 58 (at 100%)
Network size: 62
Density: 0.24
Diameter: 2
Average path length: 1.76
Average clustering coefficient: 0.85
Average degree: 14.42
Maximum degree: 61 (MR. KOUENGOUA)

Cast list

(in order of appearance)

1. MADAM PRESIDENT ♀
2. MR. HOMETOWU ♂
3. MR. KAREGYESA ♂
4. MR. KOUENGOUA ♂
5. MR. WASTLAIN ♂
6. MS. CRAWFORD ♀
7. JUDGE REDDY ♂
8. THE ENGLISH INTERPRETER
9. MR. JUMA ♂
10. WITNESS KEANE ♂
11. MR. KAREGYSA ♂
12. WITNESS TAC ♂
13. JUDGE EGOROV ♂
14. WITNESS TAK ♂
15. MR. KAREGESYA ♂
16. MS. MOBBERLEY ♀
17. WITNESS TAS ♀
18. THE KINYARWANDA INTERPRETER
19. WITNESS TAP ♀
20. WITNESS TAW ♂
21. WITNESS TBK ♂
22. WITNESS TBH ♂
23. WITNESS DES FORGES ♀



ICTR-2001-66-I

v. Athanase Seromba

📄 [gena000005](#) 📅 2001

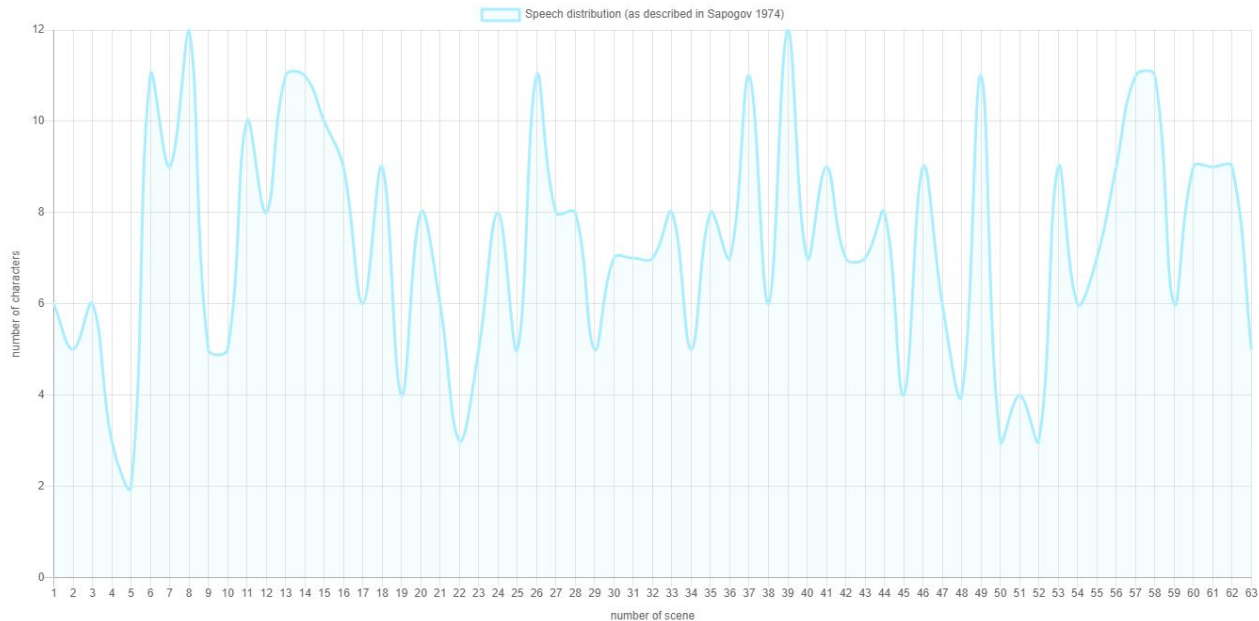
Geno DraCor



International Criminal Tribunal for Rwanda

🇷🇼 [Q478358](#)

Network [Speech distribution](#) Full text Downloads



This tab shows different ways of visualising speech distribution.

- ☒ Sapogov 1974
- ☐ Yarkho 1997 (ru), 2019 (en)
- ☐ Trilcke/Fischer et al. 2017

3. Outlining possible research applications

Getting metrics from the DraCor API

- **Basic network values**

'size', 'averageClustering', 'density', 'averagePathLength', 'maxDegreeIds',
'averageDegree', 'diameter', 'maxDegree', 'numConnectedComponents',
'numEdges'

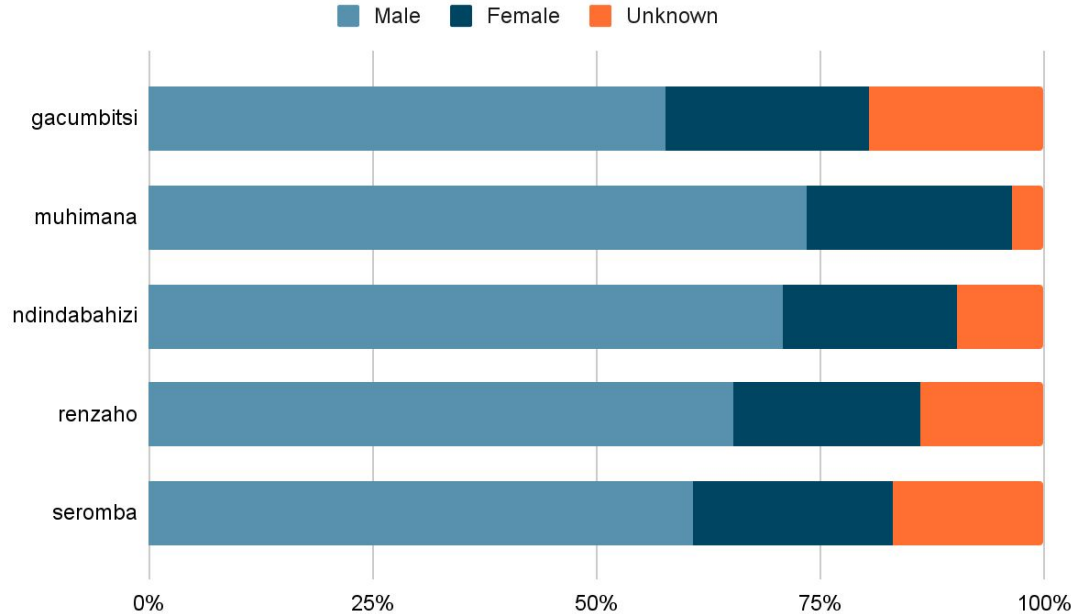
- **Speech distribution**

'numOfSpeakers', 'numOfSpeakersFemale', 'numOfSpeakersMale',
'numOfSpeakersUnknown', 'numOfPersonGroups', 'wordCountText',
'wordCountSp', 'wordCountStage'

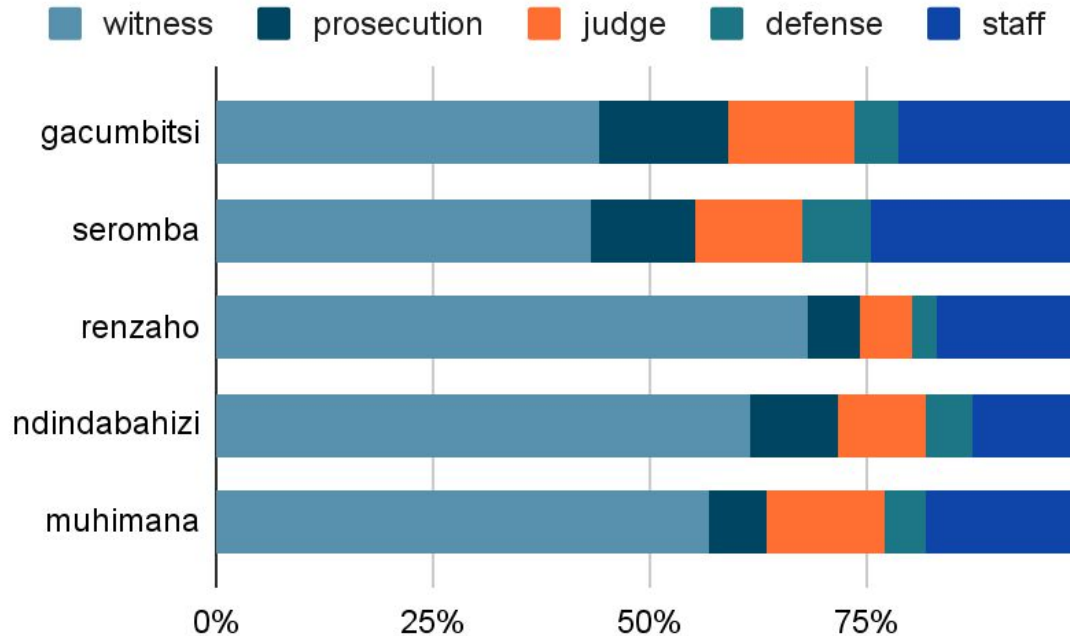
Extracting language according to gender...

Sample API call:

<https://localhost:8088/api/corpora/geno/play/rwa-muhimana/spoken-text?gender=FEMALE>



... and the characters' role in the proceedings



Sample API call:

<https://localhost:8088/api/corpora/geno/play/rwa-muhimana/spoken-text?role=judge>

Possible lines of enquiry

- **‘Distant reading’ trials through networks**
 - identify ‘core’ and ‘bridge’ characters: do their network position reflect their role within the trial? (e.g. key witnesses)
 - are there **subclusters** in correspondence of formal or informal trial phases?
- **Applying NLP techniques to extracted speeches**
 - e.g. topic modelling and sentiment analysis
 - is there any significant difference between opposing parties/between genders?
- **Exploring the theatrical dimension of judiciary proceedings**
 - comparing Genocor metadata with those from DraCor corpora
 - presence of common ‘dramatic’ structures?

Future work

- **GenoCor is still a prototype!**
 - from proof-of-concept product to open-access database
 - overcoming Docker/API limitations (e.g. file size)
- **Streamlining the onboarding process of new texts**
 - Improving semi-automatic conversion pipelines
 - Reducing the need for manual disambiguation
- **Adding more texts, especially from the ICTY**
 - structural and contextual comparative analyses
- **Improving modelling**
 - refinement of categories

References

- Börner, Ingo, et al. 'Dockerizing DraCor: A Container-based Approach to Reproducibility in Computational Literary Studies', in *DH2023 Book of Abstracts*, Graz: Universität Graz, 2023.
- Fischer, Frank, et al., 'Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama', in *DH2019 Book of Abstracts*. Utrecht: Utrecht University, 2019.
- Giovannini, Luca, et al., 'Distributed Corpus Building in Literary Studies: The DraCor Example', in *DH2023 Book of Abstracts*, Graz: Universität Graz, 2023.
- Pianzola, Federico, et al. 'Constructing the GOLEM: Graphs and Ontologies for Literary Evolution Models', in *DH2023 Book of Abstracts*, Graz: Universität Graz, 2023.
- Schirmer, Miriam. 'Mixed Methods in der Genozidforschung', in *DHd2023 Book of Abstracts*, Trier; Belval: Universität Trier und Luxemburg, 2023.

Funding

In the context of CLS INFRA, the DraCor project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101004984.



Thanks for your attention!

↻ plu.sh/genocor

✉ giovannini@uni-potsdam.de

✉ miriam.schirmer@tum.de

