
The Best of Three Worlds: Mutual Enhancement of Corpora of Dramatic Texts

— GerDraCor, German Text Archive,
TextGrid Repository —

Frank Fischer ~ Susanne Haaf ~ Marius Hug

Background

- Corpora of dramatic texts
 - TextGrid Repository
 - GerDraCor (German Drama Corpus)
 - DTA (Deutsches Textarchiv)
- Aims
 - Increase corpus sizes
 - Mutually enhance depth and quality of markup
 - Enable interoperability of corpora (per DTABf)
 - Ensure long-term availability of upgraded corpus data
- Methods
 - Implement workflows for conversion into DTABf
 - Mutually enrich all corpora in question
 - Integrate upgraded corpus data in CLARIN repository

Joint Effort



NATIONAL RESEARCH
UNIVERSITY



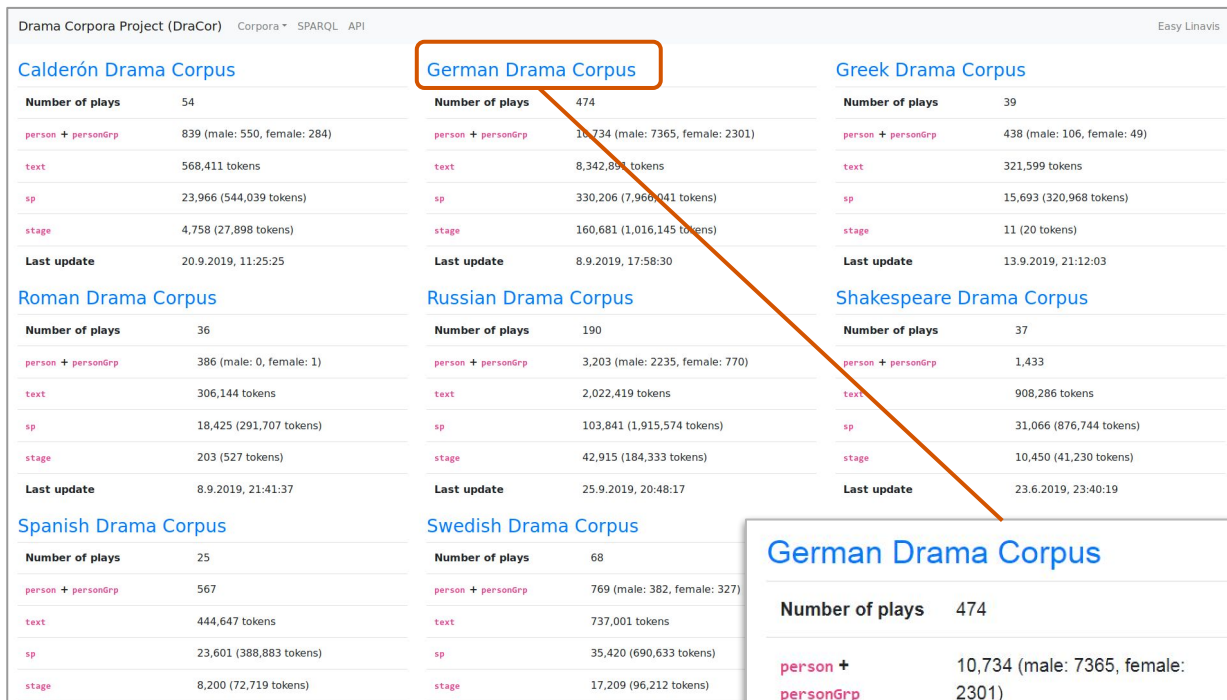
Corpora Involved

DraCor

- **Drama Corpora** platform
- dracor.org
- multilingual drama corpora
- **474 multi-sourced German plays** (mostly from TextGrid Repository)
- API ("Programmable Corpora")

Drama Corpora Project (DraCor) Corpora - SPARQL API Easy Linavis

Calderón Drama Corpus		German Drama Corpus		Greek Drama Corpus	
Number of plays	54	Number of plays	474	Number of plays	39
person + personGrp	839 (male: 550, female: 284)	person + personGrp	10,734 (male: 7365, female: 2301)	person + personGrp	438 (male: 106, female: 49)
text	568,411 tokens	text	8,342,891 tokens	text	321,599 tokens
sp	23,966 (544,039 tokens)	sp	330,206 (7,966,041 tokens)	sp	15,693 (320,968 tokens)
stage	4,758 (27,898 tokens)	stage	160,681 (1,016,145 tokens)	stage	11 (20 tokens)
Last update	20.9.2019, 11:25:25	Last update	8.9.2019, 17:58:30	Last update	13.9.2019, 21:12:03
Roman Drama Corpus		Russian Drama Corpus		Shakespeare Drama Corpus	
Number of plays	36	Number of plays	190	Number of plays	37
person + personGrp	386 (male: 0, female: 1)	person + personGrp	3,203 (male: 2235, female: 770)	person + personGrp	1,433
text	306,144 tokens	text	2,022,419 tokens	text	908,286 tokens
sp	18,425 (291,707 tokens)	sp	103,841 (1,915,574 tokens)	sp	31,066 (876,744 tokens)
stage	203 (527 tokens)	stage	42,915 (184,333 tokens)	stage	10,450 (41,230 tokens)
Last update	8.9.2019, 21:41:37	Last update	25.9.2019, 20:48:17	Last update	23.6.2019, 23:40:19
Spanish Drama Corpus		Swedish Drama Corpus			
Number of plays	25	Number of plays	68		
person + personGrp	567	person + personGrp	769 (male: 382, female: 327)		
text	444,647 tokens	text	737,001 tokens		
sp	23,601 (388,883 tokens)	sp	35,420 (690,633 tokens)		
stage	8,200 (72,719 tokens)	stage	17,209 (96,212 tokens)		

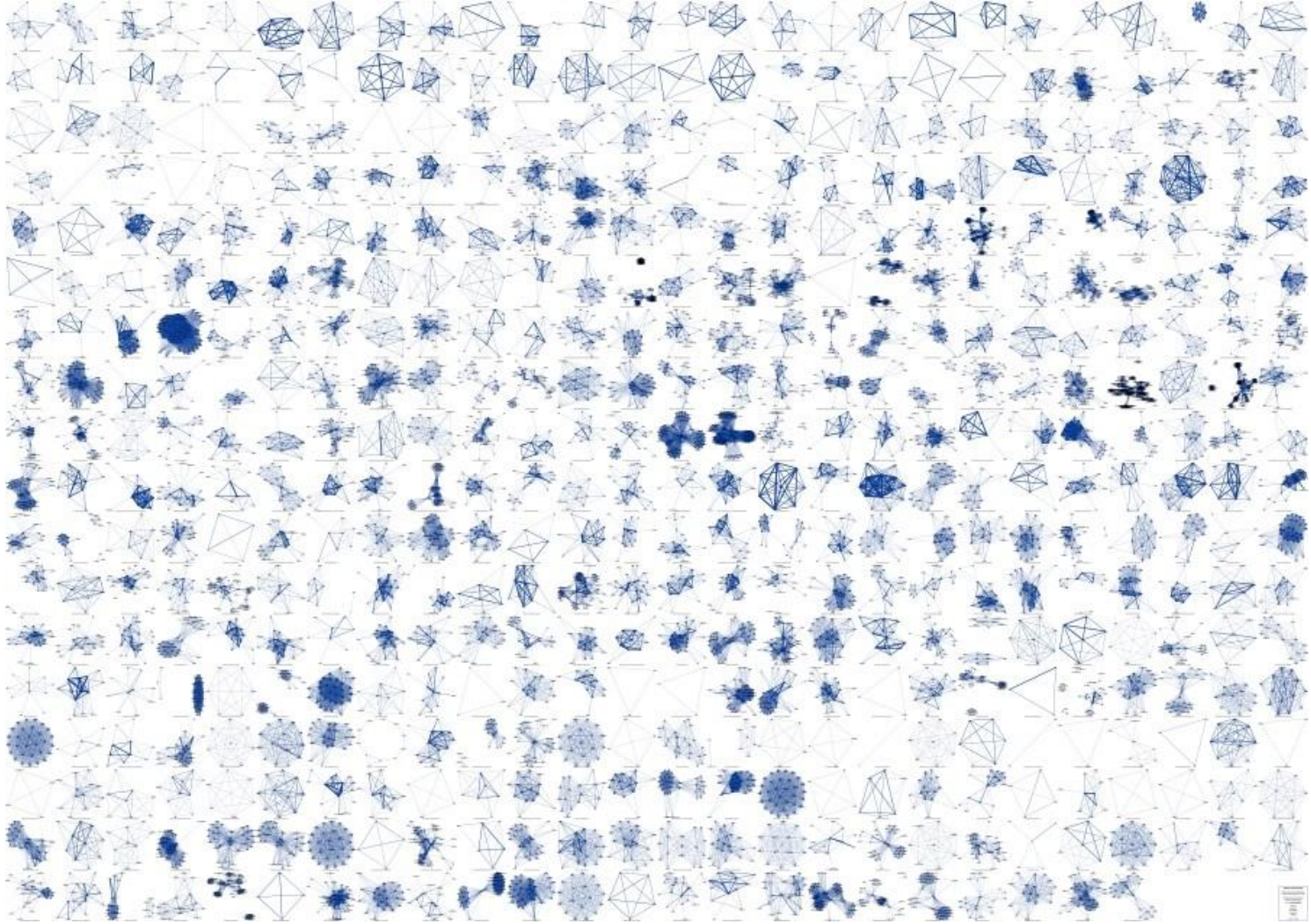


German Drama Corpus

Number of plays	474
person + personGrp	10,734 (male: 7365, female: 2301)
text	8,342,891 tokens
sp	330,206 (7,966,041 tokens)
stage	160,681 (1,016,145 tokens)
Last update	8.9.2019, 17:58:30

Distant-Reading Showcase (2016)

- social networks
- extracted from 465 plays
- chronological order



Deutsches Textarchiv (DTA)

- Corpus of historical New High German texts
- Mostly printed texts
- Growing number
- Currently 5,961 docs.
 - ~1 million pages
 - ~ 4,000 volumes of scientific, functional or fictional literature;
 - ~ 2,000 newspaper issues
 - **101 plays**

Anmelden (DTAQ)

DTA [Hilfe](#)

in den Titeldaten im Korpus in der Dokumentation



Kleist, Heinrich von: Die Schlacht bei Fehrbellin. Berlin, 1822.

BIBLIOGRAPHISCHE ANGABEN

URN:	urn:nbn:de:kobv:b4-200905193140
Titel:	Die Schlacht bei Fehrbellin
Untertitel:	Schauspiel in fünf Akten
weiterer Titel:	Prinz Friedrich von Homburg
Autor/in:	Heinrich von Kleist (GND, Wikipedia, ADB/NDB)
Erscheinungsjahr:	1822
Verlag/Drucker:	Wallishäuser; Reimer
Ort:	Berlin
Auflage:	1. Auflage
Bibliothek:	Staatsbibliothek zu Berlin – Preußischer Kulturbesitz
Signatur:	SBB-PK, 348074 R

INFORMATIONEN ZUM WERK

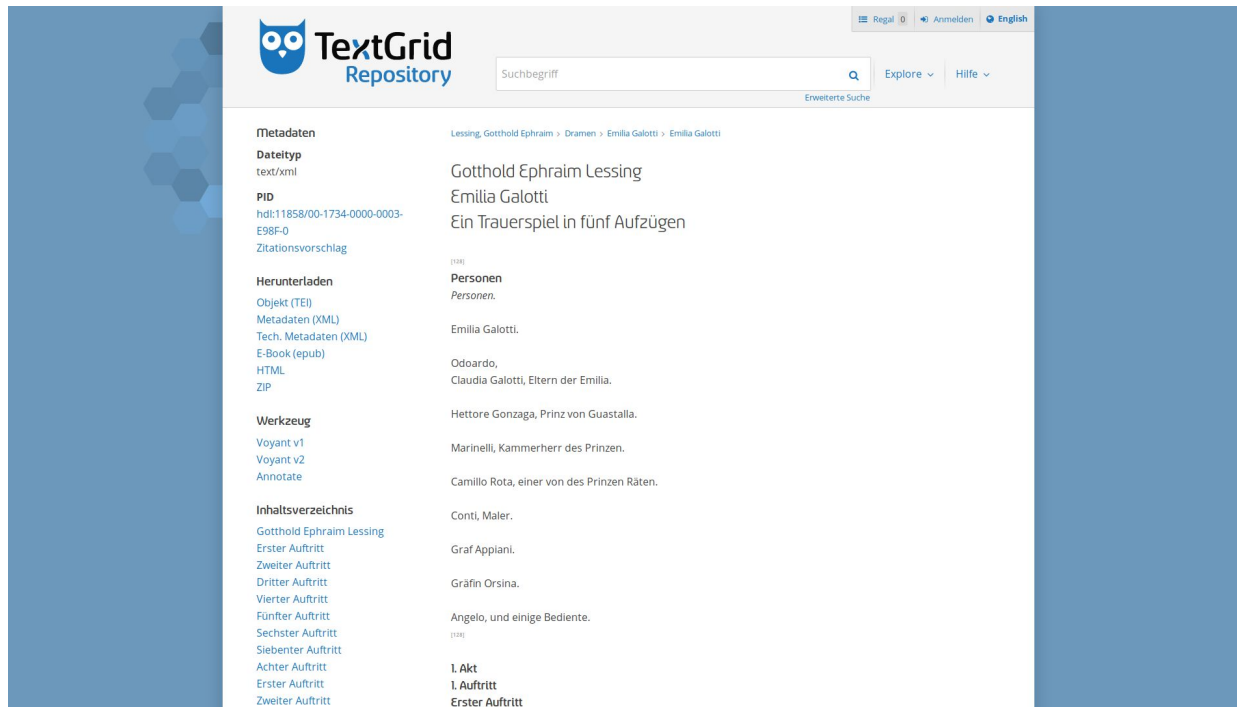
Publikationstyp:	Monographie
Verfügbarkeit:	Text (TEI-XML-, HTML-, TCF-, E-Book-Fassung): CC BY-NC 3.0 Weitere Informationen: Nutzungsbedingungen.
Schriftart:	Fraktur
Genre:	Belletristik :: Drama
im DTA seit:	2008-05-21 14:44:44
Korpus:	DTA-Kernkorpus

Berlin SBB-PK, 348074 R

[Text-Bild-Ansicht öffnen ...](#)

TextGrid Repository (TGR)

- freely usable repository
- contains "Die Digitale Bibliothek" (= thousands of texts of all kinds, among which are **666 plays**)



The screenshot displays the TextGrid Repository interface. At the top, there is a navigation bar with the TextGrid logo (an owl) and the text 'TextGrid Repository'. To the right of the logo is a search bar with the placeholder text 'Suchbegriff' and a search icon. Further right are links for 'Regal 0', 'Anmelden', and 'English'. Below the search bar is a dropdown menu with 'Explore' and 'Hilfe' options, and a link for 'Erweiterte Suche'.

The main content area is divided into two columns. The left column contains metadata and download options:

- Metadaten**
- Datentyp**: text/xml
- PID**: hdl:11858/00-1734-0000-0003-E98F-0
- Zitationsvorschlag**
- Herunterladen**: Objekt (TEI), Metadaten (XML), Tech. Metadaten (XML), E-Book (epub), HTML, ZIP
- Werkzeug**: Voyant v1, Voyant v2, Annotate
- Inhaltsverzeichnis**: Gotthold Ephraim Lessing, Erster Auftritt, Zweiter Auftritt, Dritter Auftritt, Vierter Auftritt, Fünfter Auftritt, Sechster Auftritt, Siebenter Auftritt, Achter Auftritt, Erster Auftritt, Zweiter Auftritt

The right column shows the title and author information:

- Lessing, Gotthold Ephraim > Dramen > Emilia Galotti > Emilia Galotti
- Gotthold Ephraim Lessing
- Emilia Galotti
- Ein Trauerspiel in fünf Aufzügen

Below the title, there are sections for 'Personen' and 'I. Akt':

- Personen**: Personen, Emilia Galotti, Odoardo, Claudia Galotti, Eltern der Emilia, Hettore Gonzaga, Prinz von Guastalla, Marinelli, Kammerherr des Prinzen, Camillo Rota, einer von des Prinzen Räten, Conti, Maler, Graf Appiani, Gräfin Orsina, Angelo, und einige Bediente.
- I. Akt**: I. Auftritt, Erster Auftritt

Three Corpora

- GerDraCor
 - *Strength*: actively maintained and extended; carefully assigned speaker-IDs; repaired markup flaws of TextGridRep
 - *Weakness*: corpus not balanced enough yet
 - *Benefits from*: granularity of DTA annotations
- TextGridRep
 - *Strength*: create and download individually created corpora
 - *Weakness*:
 - minor text loss due to conversion issues
 - problems with <stage>- and <title>-annotations
 - *Benefits from*: DTA-XML-versions (via CAB), GDC stage-annotations, <title> annotations

Three Corpora

- DTA/DTAQ:
 - *Strength:*
 - Accuracy
 - Download in various formats (TEI, txt, TCF, HTML, ...)
 - Analysis tools directly on platform
 - Collaborative quality assurance platform DTAQ
 - Text in various processing stages (original, modernized, lemmatized, ...)
 - *Weakness:* semi-automatically applied speaker-IDs
 - *Benefits from:* GerDraCor speaker-IDs

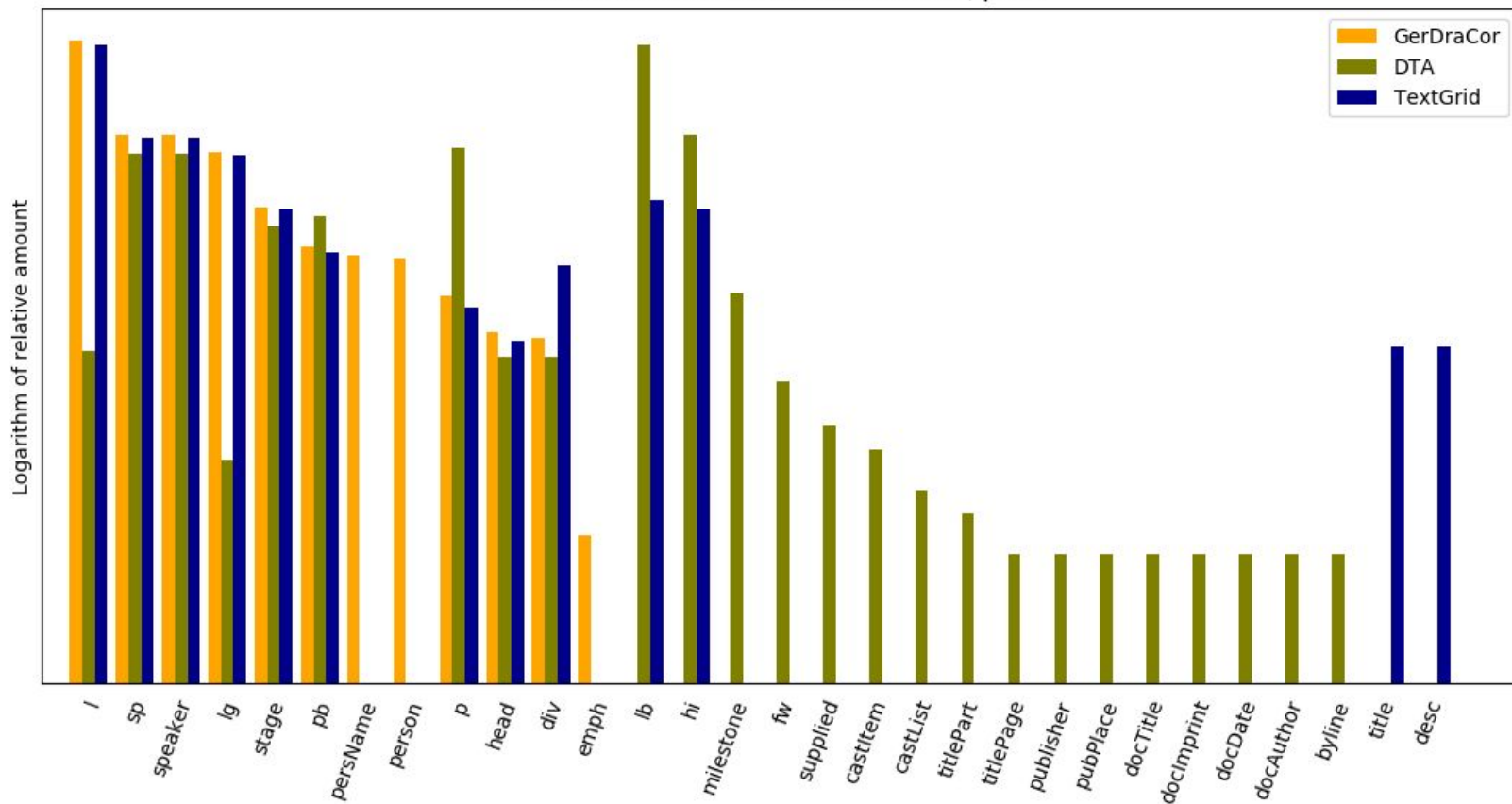
Harmonization of Resources

Differences of Text and Tagging

- Three corpora of TEI texts → three versions of transcriptions and TEI encoding
- Transcription:
 - Differentiation between f and s, uͤ and ü, ... (DTA) vs. normalized writing of s, ü, ... (GerDraCor)
- TEI Encoding
 - <hi> (DTA, TextGridRep) vs. <emph> (GerDraCor)
 - <titlePage> (DTA); <lb> (DTA, TextGridRep); <person> (GerDraCor)
 - Manually (GerDraCor) vs. semi-automatically (DTA) applied speaker IDs

Differences in TEI Encoding

TEI elements in Goethe's drama Faust, part 1



Merging

Example Drama

Collin: *Coriolan*

```
<text>
  <front>
    <pb n="1"/>
    <titlePage type="main">
      <docTitle>
        <titlePart type="main">Coriolan.</titlePart>
        <titlePart type="sub">Ein Trauerspiel in fünf Aufzügen.</titlePart>
      </docTitle>
      <byline>von<lb/>
        <docAuthor>Collin.</docAuthor></byline>
      <docImprint><pubPlace>Berlin,</pubPlace><lb/> bei <publisher>Johann Friedrich
        Unger.</publisher><lb/>
        <docDate>1804.</docDate></docImprint>
    </titlePage>
    <pb n="2"/>
    <pb n="3"/>
    <castList>
      <head><emph>Personen.</emph></head>
      <p><emph>Römer.</emph></p>
      <castItem><emph>Cajus Marcius Coriolanus.</emph></castItem>
      <castItem><emph>Veturia</emph>, seine Mutter.</castItem>
```

Titlepage from DTA

Merging

Ex. Drama

Collin: *Coriolan*

Verse encoding from
GerDraCor

```
<l>Verzehret mich. Ach! solls denn ewig wahren?</l>
</lg>
</sp>
<sp who="#veturia">
  <speaker>Veturia.</speaker>
  <l part="I">Die Sonne sinket -</l>
</sp>
<sp who="#volumnia">
  <speaker>Volumnia.</speaker>
  <lg>
    <l part="F">Nun, so wird es bald</l>
    <l part="I">Sich enden.</l>
  </lg>
</sp>
<sp who="#veturia">
  <speaker>Veturia.</speaker>
  <lg>
    <l part="F">Denke dir das Schlimmste, Tochter!</l>
    <pb n="8"/>
```

Merging

Example Drama
Collin: Coriolan

```
<l>Verzehret mich. Ach! solls denn ewig wahren?</l>
```

```
</lg>
```

```
</sp>
```

```
<sp who="#veturia">
```

```
<speaker>Veturia.</speaker>
```

```
<l part="I">Die Sonne sinket -</l>
```

```
</sp>
```

```
<sp who="#volumnia">
```

```
<speaker>Volumnia.</speaker>
```

```
<lg>
```

```
<l part="F">Nun, so wird es bald</l>
```

```
<l part="I">Sich enden.</l>
```

```
</lg>
```

```
</sp>
```

```
<sp who="#veturia">
```

```
<speaker>Veturia.</speaker>
```

```
<lg>
```

```
<l part="F">Denke dir das Schlimmste, Teufel!</l>
```

```
<pb n="8"/>
```

Verse encoding from
GerDraCor

Verse encoding in
DTA

```
<sp who="#VET">
```

```
<speaker><hi rendition="#g">Veturia</hi>.</speaker><lb/>
```

```
<p>Die Sonne &#x017F;inket &#x2014;</p>
```

```
</sp><lb/>
```

```
<sp who="#VOLU">
```

```
<speaker><hi rendition="#g">Volumnia</hi>.</speaker><lb/>
```

```
<p><hi rendition="#et">Nun, &#x017F;o wird es bald</hi><lb/>
```

```
Sich enden.</p>
```

```
</sp><lb/>
```

```
<sp who="#VET">
```

```
<speaker><hi rendition="#g">Veturia</hi>.</speaker><lb/>
```

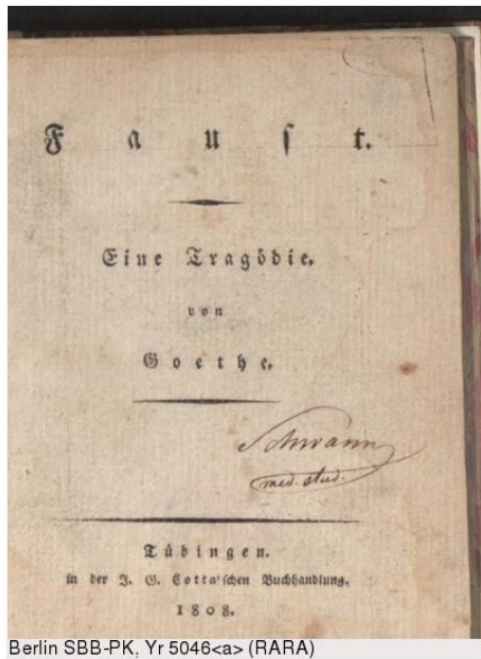

Benefits

Benefits

- Analysis options on different platforms for similarly tagged texts
- Interoperability of resources across projects
 - Additional drama texts of TextGridRep and GerDraCor become analyzable with DTA tools
 - Additional drama texts of DTA become analyzable with GerDraCor tools
- Drama corpus via CLARIN repository @BBAW (long-term preservation)

Example: Analysis @DTA (1/4)

Goethe, Johann Wolfgang von: Faust. Eine Tragödie. Tübingen, 1808.



Berlin SBB-PK, Yr 5046<a> (RARA)

Cover in hoher Auflösung

Informationen


Quelle:	CN
Publikationstyp:	Monographie
Umfang:	321 Scans ca. 201837 Zeichen ca. 30773 Wortformen ca. 7112 Oberflächentypes
Schriftart:	Fraktur
Genre:	Belletristik :: Drama
im DTA seit:	2008-01-24 11:25:58
zuletzt geändert:	2017-08-10 15:46:58
Verfügbarkeit:	Text (TEI-XML-, HTML-, TCF-, E-Book-Fassung): CC BY-NC 3.0 Weitere Informationen: Nutzungsbedingungen.

- Anmerkung zu Informationen verfassen

Metadaten

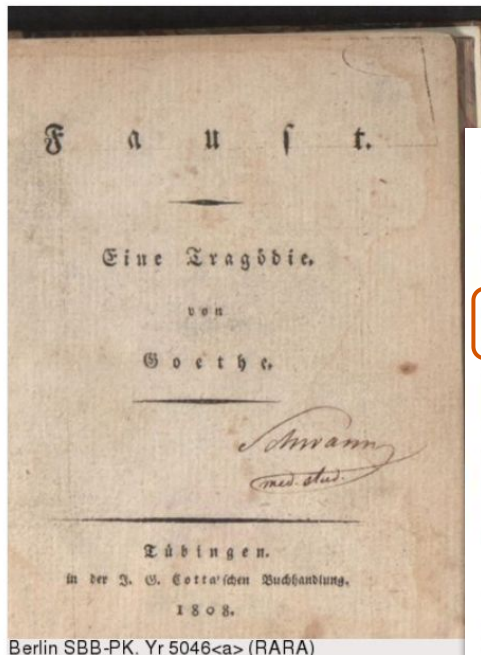
URN:	urn:nbn:de:kobv:b4-200905191726
Titel:	Faust
Untertitel:	Eine Tragödie

Ansichten

- Korrekturumgebung
- Webversion
- Faksimiles (DFG-Viewer)
- Download:
 - Text (UTF-8, Seitenumbrüche als ASCII \014)
 - TEI/XML (mit Silbentrennung)
 - HTML (mit Silbentrennung)
 - TCF (text annotation layer)
 - TCF (tokenisiert, serialisiert, lemmatisiert, normalisiert)
 - TEI/XML (inkl. att.linguistic)
 - TEI-Header
 - CMDI
 - Dublin Core
- Lemmata:
 - nach Frequenz
 - nach Frequenz (nur Nomen)
 - Wortwolke
 - Wortwolke (nur Nomen)
- Wortformen (Types):
 - nach Frequenz
 - nach Frequenz (nur Nomen)
 - Wortwolke
 - Wortwolke (nur Nomen)
- Voyant Tools ?:
 - transliterierter Text
 - normalisierter Text

Example: Analysis @DTA (2/4)

Goethe, Johann Wolfgang von: Faust. Eine Tragödie. Tübingen, 1808.



Berlin SBB-PK, Yr 5046<a> (RARA)

Cover in hoher Auflösung

Informationen

Quelle: CN
Publikationstyp: Monographie

Faust	256
Herr	79
Geist	78
Herz	58
Welt	55
Tag	50
Zeit	44

URN: urn:nbn:de:koebv:b4-20090519
Titel: Faust
Untertitel: Eine Tragödie

Ansichten

- Korrekturumgebung
- Webversion
- Faksimiles (DFG-Viewer)
- Download:
 - Text (UTF-8, Seitenumbrüche als ASCII \014)
 - TEI/XML (mit Silbentrennung)
 - HTML (mit Silbentrennung)
 - TCF (text annotation layer)
 - TCF (tokenisiert, serialisiert, lemmatisiert, normalisiert)
 - TEI/XML (inkl. att.linguistic)
 - TEI-Header

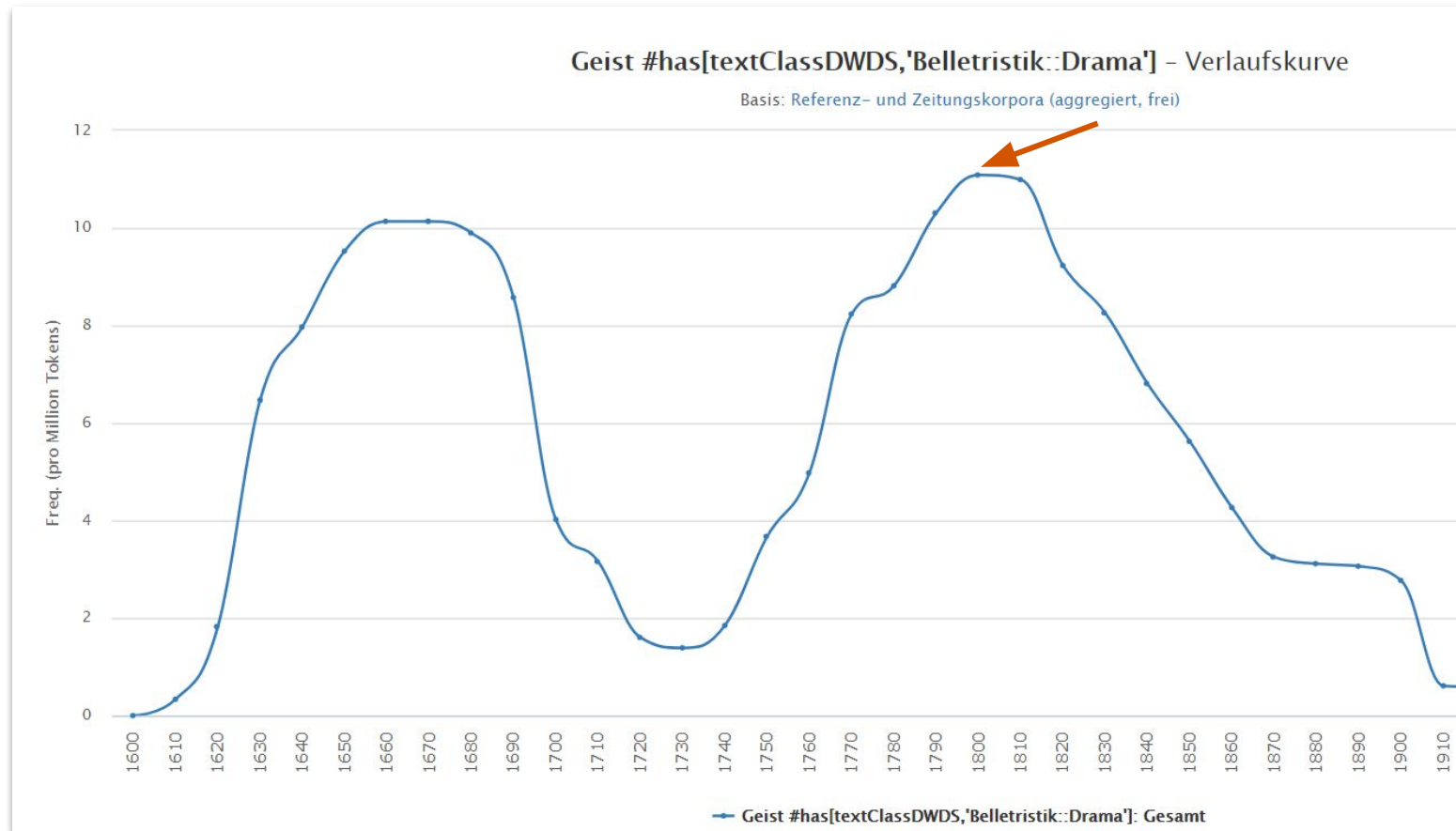
• Lemmata:

- nach Frequenz
- nach Frequenz (nur Nomen)
- Wortwolke
- Wortwolke (nur Nomen)

• Wortformen (Types):

- nach Frequenz
- nach Frequenz (nur Nomen)
- Wortwolke
- Wortwolke (nur Nomen)

Example: Analysis @DTA (3/4)



Example: Analysis @GerDraCor (1/2)

Research-driven platform, papers presented at recent DH conferences:

- conditions for a network analysis of dramatic texts ([DH2015](#))
- small-world phenomenon in drama ([DH2016](#))
- progressive structuration of drama networks/„plot“ ([DH2017](#))
- catching protagonists: typology of characters ([DH2018](#))

Example: Analysis @GerDraCor (2/2)

Shiny Dracor

Choose a corpus
German Drama Corpus

Choose a writer from a list:
Kleist, Heinrich von

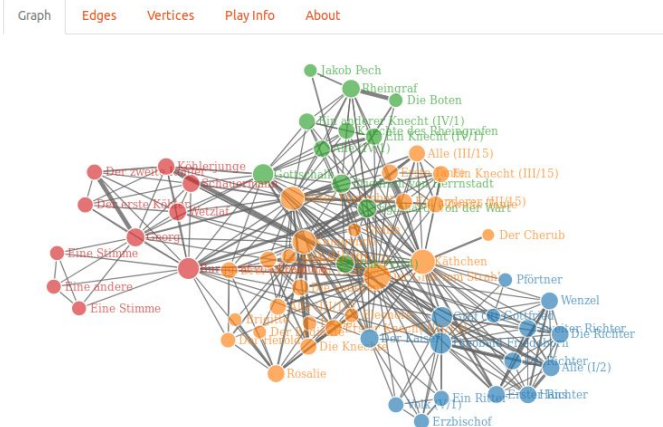
Choose his/her play to visualize:
Das Käthchen von Heilbronn oder die Feuerprobe

Choose a metric for nodes size:
Degree

Choose clusterization algorithm:
cluster_optimal

Select charge: 6.15 Select font size: 12

Nodes size: 1 Edges size: 1



DraCor Shiny App:

- <https://shiny.dracor.org/>
- network analysis of character interactions
- data received via DraCor API

Perspectives

Perspectives

- TGR → GerDraCor
 - Finalized ([documentation](#))
- GerDraCor ↔ DTA
 - Workflow built up
 - Conversion continuing
 - Expected finalization: 2020
- DTA ↔ TGR
 - Working package within CLARIAH-DE (03/2019–03/2021)
 - Conversion of selected but substantial TGR-resources into DTABf
 - In this course (manually) correcting and (automatically) enriching metadata
 - Integrate TGR into DTA- and CLARIN-infrastructure

Thank you!

Frank Fischer (Higher School of Economics, Moscow · DARIAH-EU)

frank.fischer@dariah.eu

Susanne Haaf (CLARIN-D at BBAW, Berlin)

haaf@bbaw.de

Marius Hug (CLARIAH-DE at BBAW, Berlin)

marius.hug@bbaw.de