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

GerDraCor, German Text Archive, TextGrid Repository

Frank Fischer ~ Susanne Haaf ~ Marius Hug

CLARIN Annual Conference 2019 · Leipzig, 2 Oct 2019

CLARIN

Background

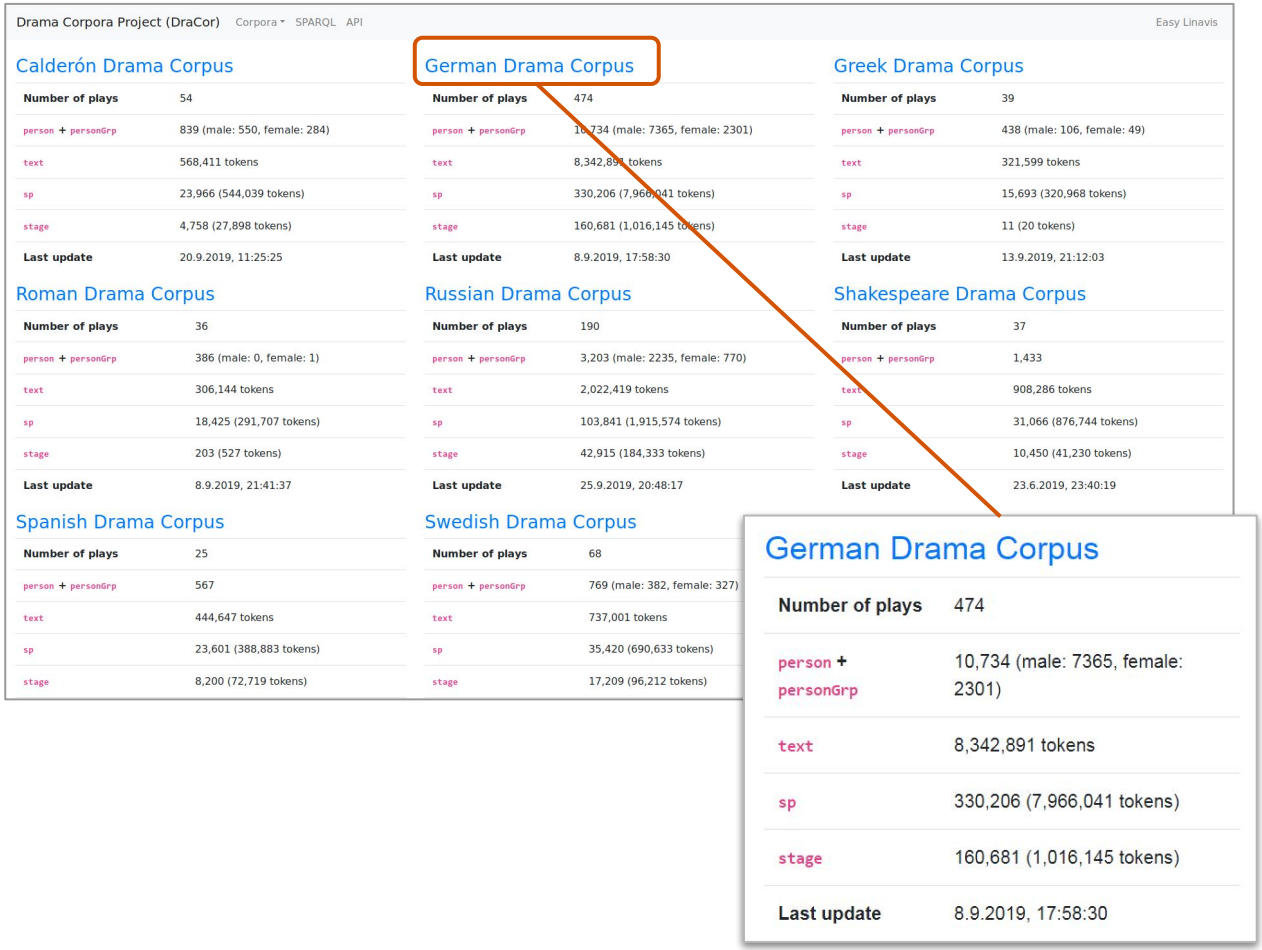
- Corpora of dramatic texts
  - TextGrid Repository
  - GerDraCor (German Drama Corpus)
  - DTA (Deutsches Textarchiv)
- Aims
  - Increase corpus sizes
  - Mutually enhance depth and quality of markup
  - Enable interoperability of corpora (per DTABf)
  - Ensure long-term availability of upgraded corpus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

Joint Effort

Corpora Involved

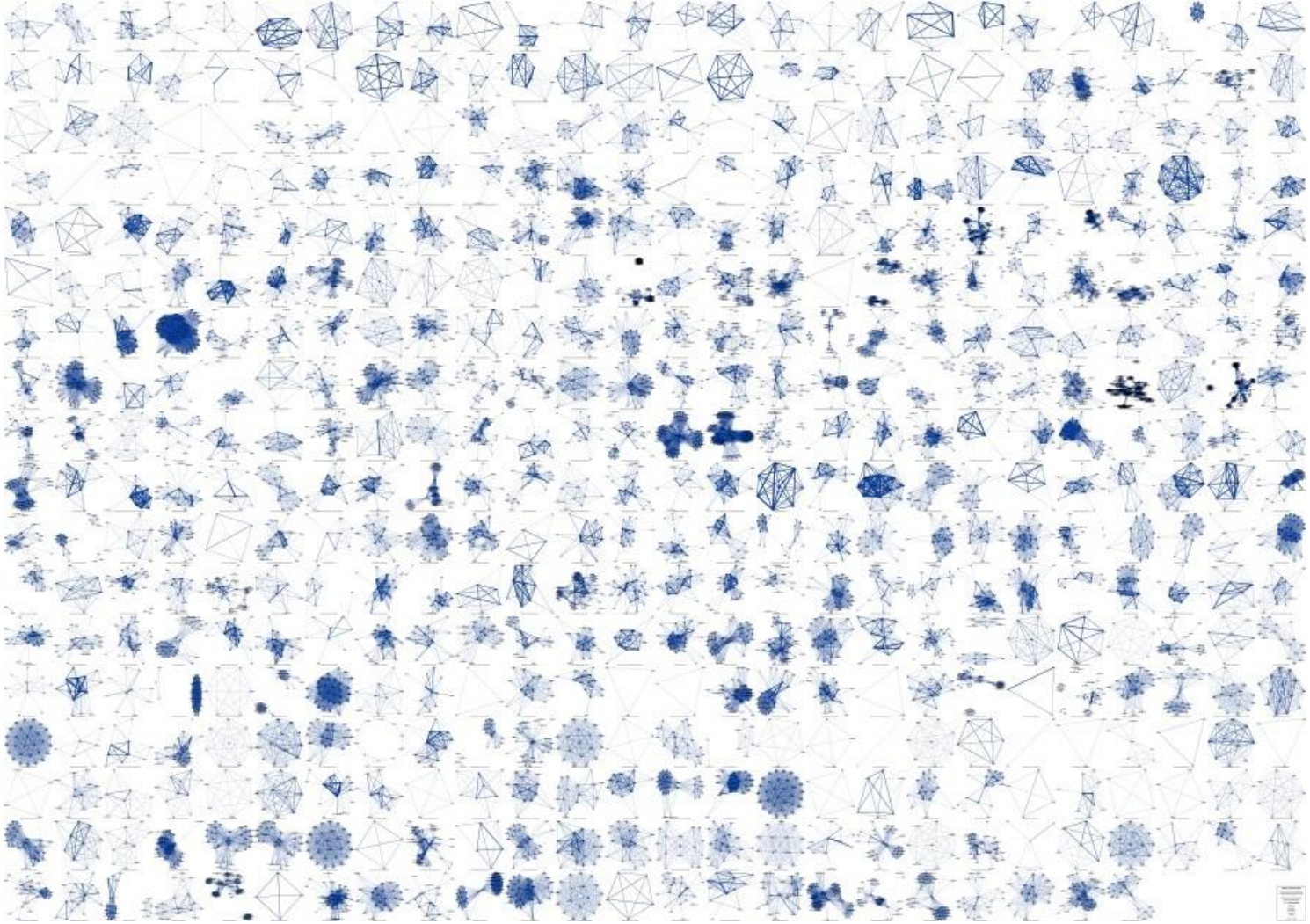
DraCor

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



Distant-Reading Showcase (2016)

- social networks
- extracted from 465 plays
- chronological order

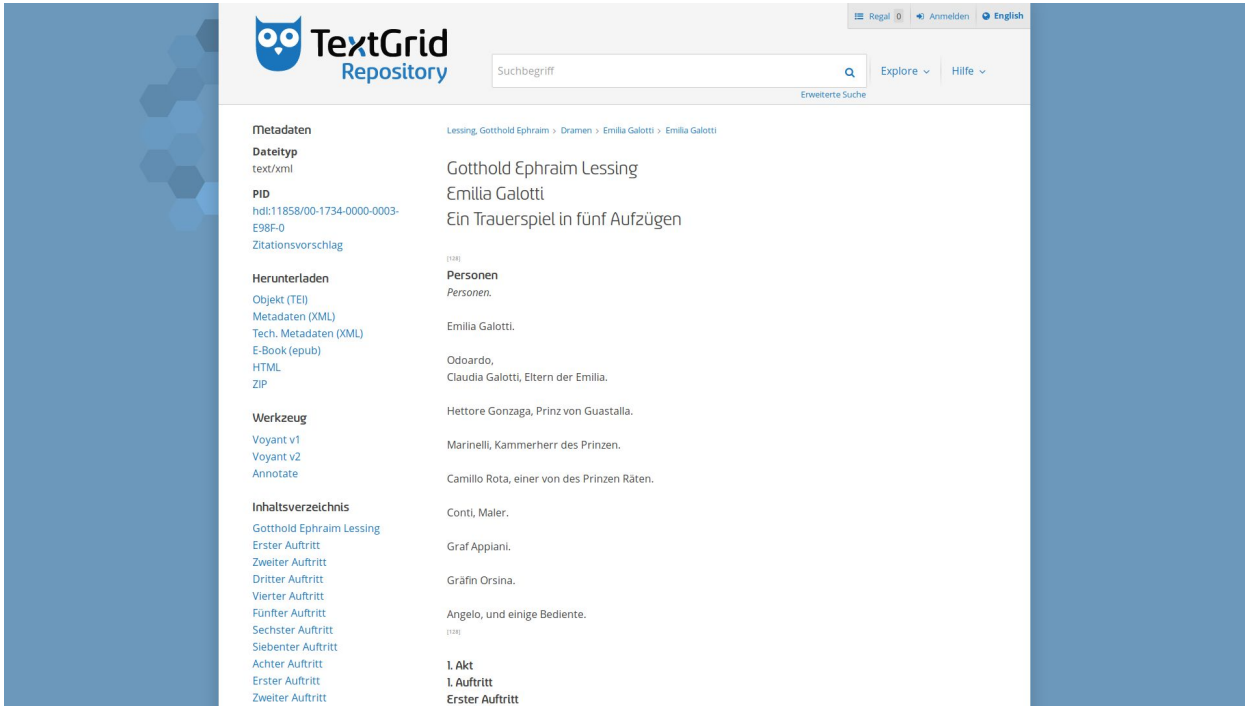


Deutsches Textarchiv (DTA)

- 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**

TextGrid Repository (TGR)

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



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

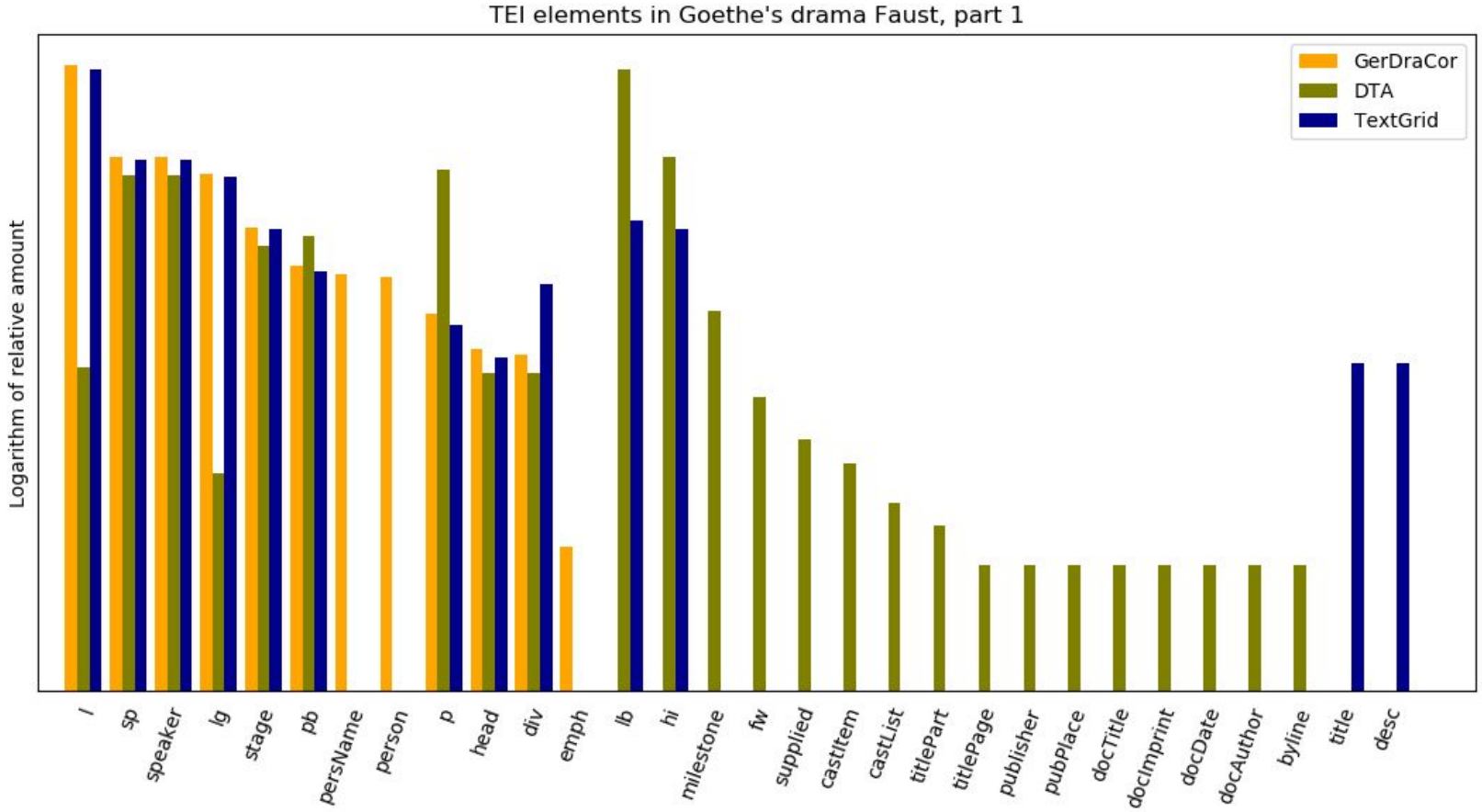
- DTA/DTAQ:
  - *Strength*:
    - Accuracy
    - Download in various formats (TEI, txt, TCF, HTML, ...)
    - Analysis tools directly on platform
    - Collaborative quality assurance platform DTAQ
    - Text in various processing stages (original, modernized, lemmatized, ...)
  - *Weakness*: semi-automatically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



Merging

Example Drama

Collin: *Coriolan*

Titlepage from DTA

```
<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>
```

Merging

Ex. Drama
Collin: *Coriolan*

Verse encoding from GerDraCor

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

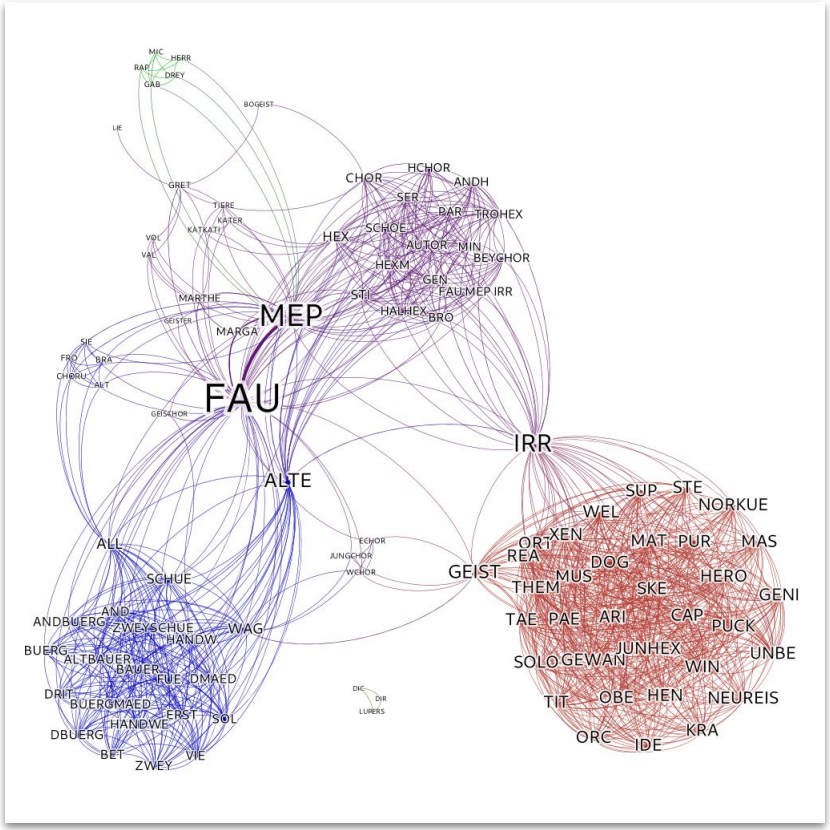
Verse encoding from GerDraCor

Verse encoding in DTA

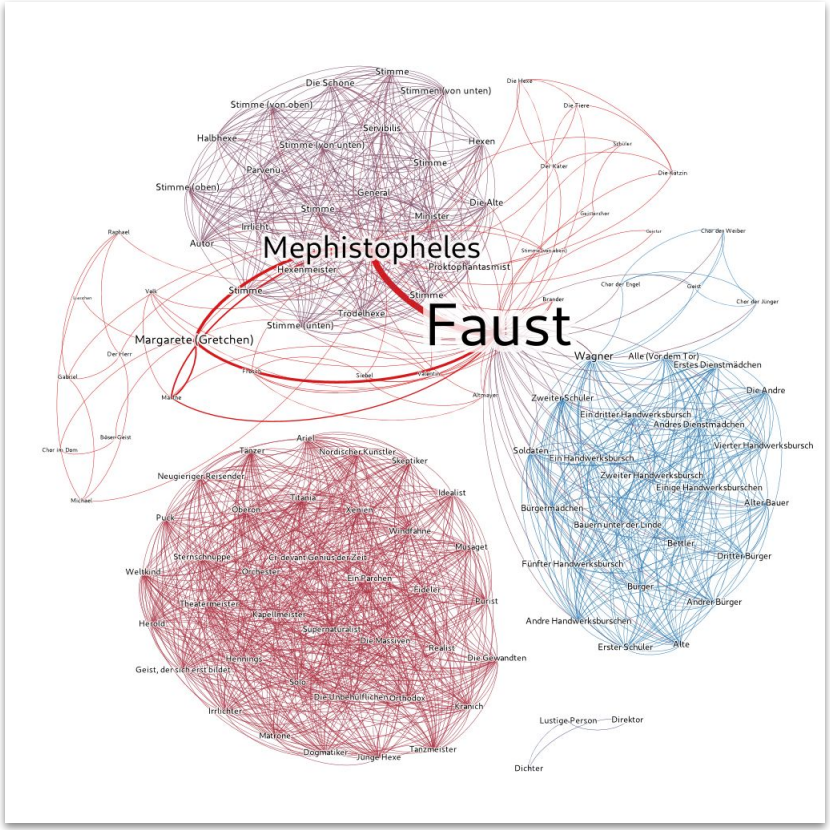
```
<sp who="#VET">
  <speaker><hi rendition="#g">Veturia</hi>.</speaker><lb/>
  <p>Die Sonne &#x017F;inket &#x2014;</p>
</sp><lb/>
<sp who="#VOLU">
  <speaker><hi rendition="#g">Volumnia</hi>.</speaker><lb/>
  <p><hi rendition="#et">Nun, &#x017F;o wird es bald</hi><lb/>
    Sich enden.</p>
</sp><lb/>
<sp who="#VET">
  <speaker><hi rendition="#g">Veturia</hi>.</speaker><lb/>
```

Differences in Speaker IDs

Ex. Drama Goethe: Faust



Extracted from DTA version (104 nodes)



Extracted from GerDraCor version (115 nodes)

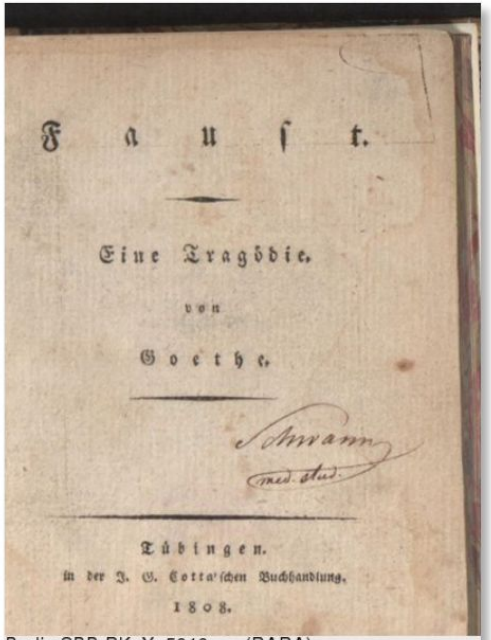
Benefits

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 `/[[:alnum:]]+/`
	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

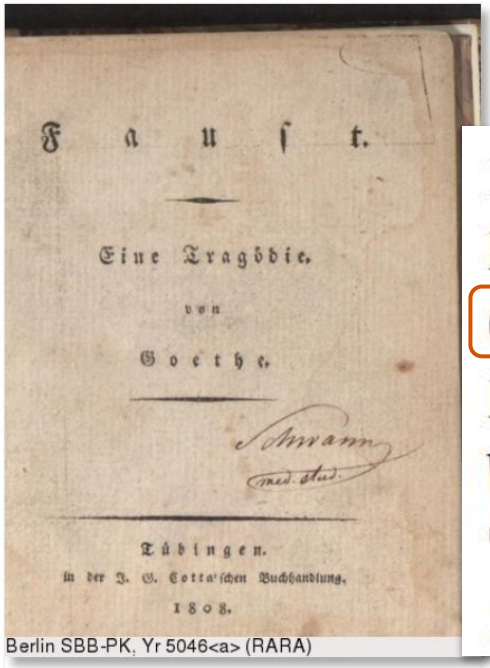
URN:	urn:nbn:de:kobv:b4-200905191726
Titel:	Faust
Untertitel:	Eine Tragödie

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

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

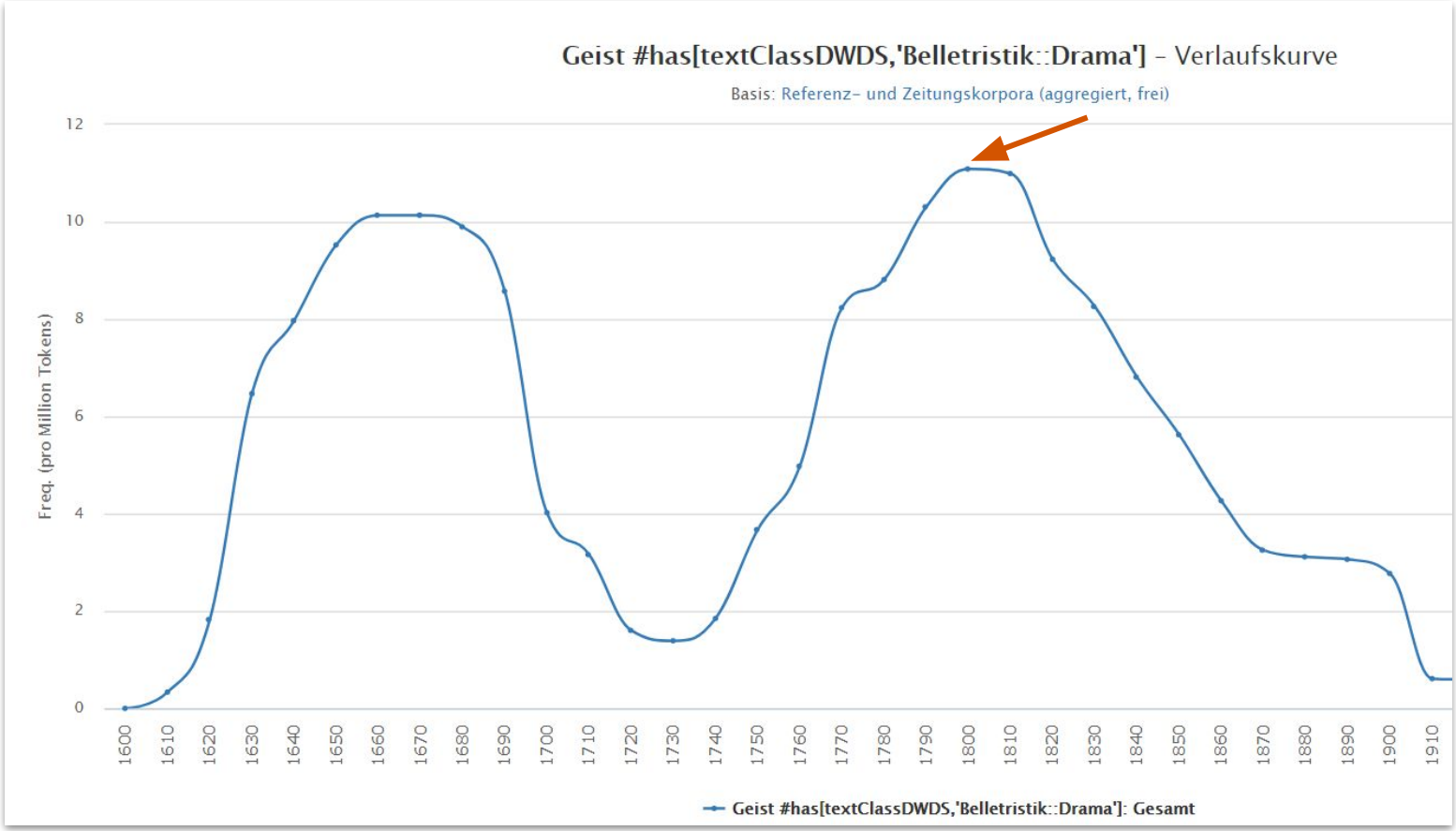
URN: urn:nbn:de:kobv:b4-2009051
Titel: Faust
Untertitel: Eine Tragödie

Ansichten

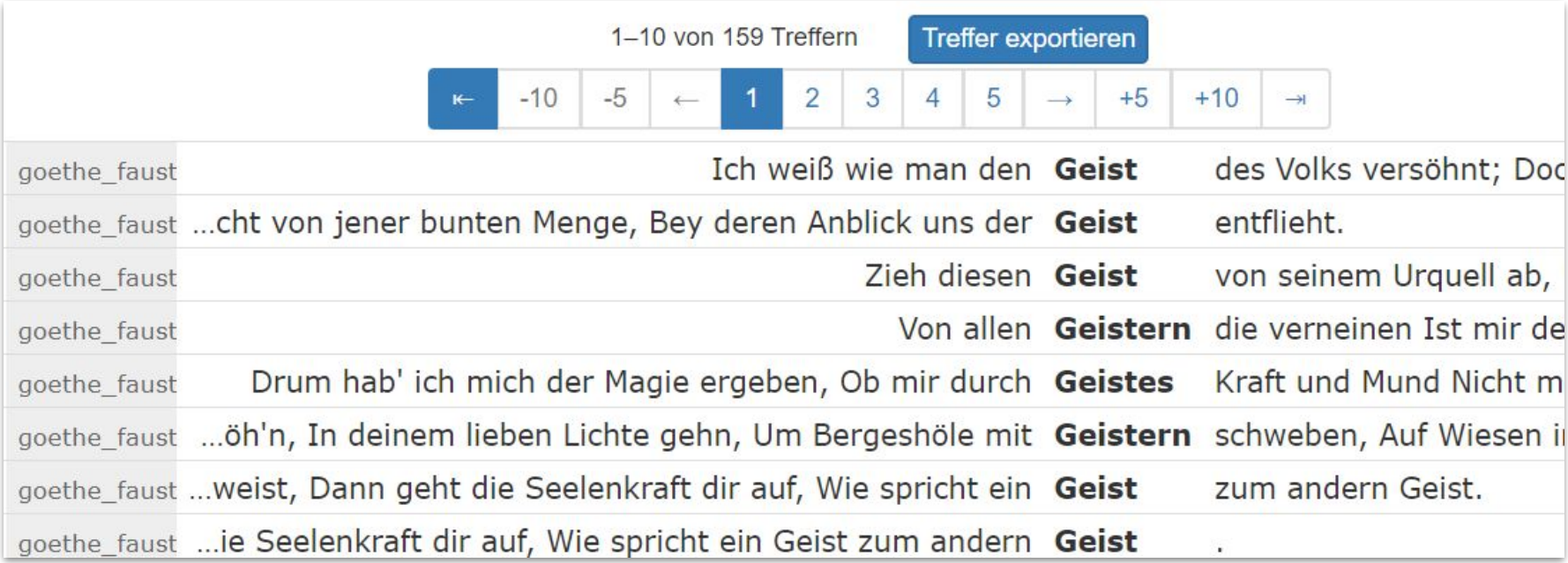
- Korrekturumgebung
- Webversion
- Faksimiles (DFG-Viewer)
- Download:
  - Text (UTF-8, Seitenumbrüche als ASCII `\014`)
  - TEI/XML (mit Silbentrennung)
  - HTML (mit Silbentrennung)
  - TCF (text annotation layer)
  - TCF (tokenisiert, serialisiert, lemmatisiert, normalisiert)
  - TEI/XML (inkl. att.linguistic)
  - 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 @DTA (4/4)

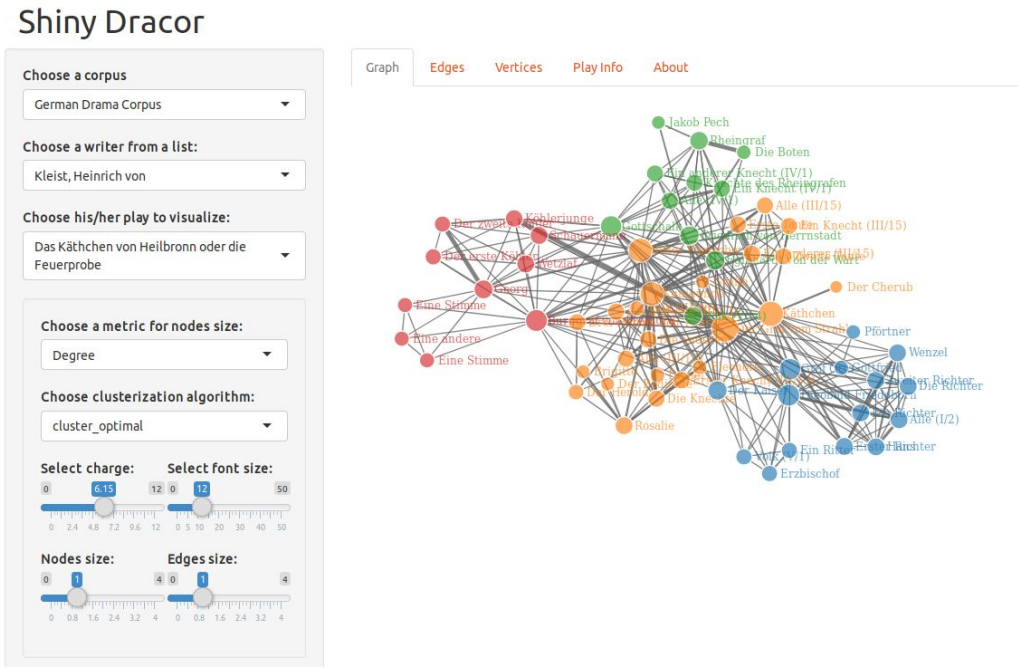


Example: Analysis @GerDraCor (1/2)

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)



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