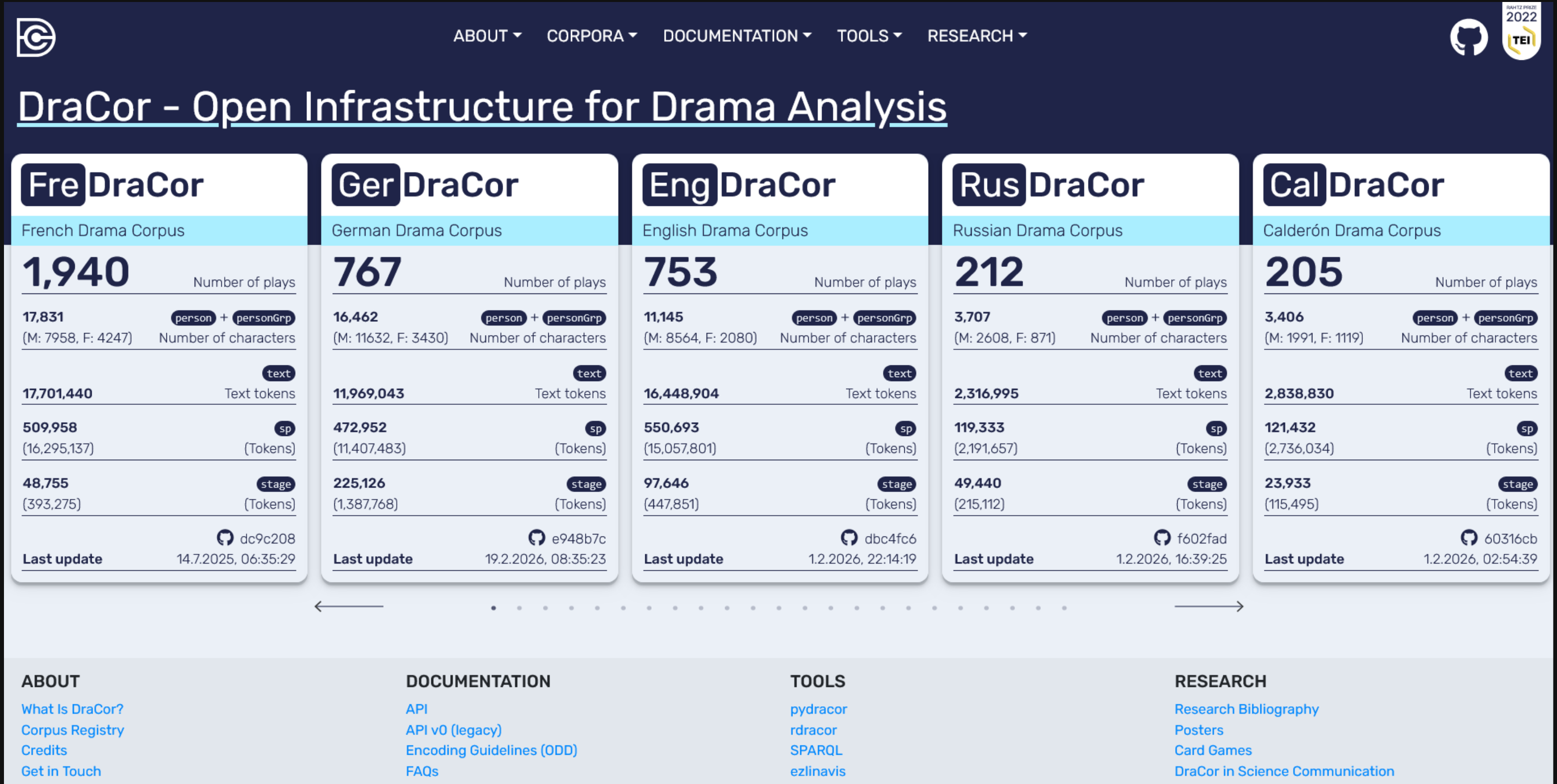
Meet DraCor

An Introduction and a Research Showcase

Luca Giovannini
University of Potsdam, Germany

Cultural Analytics Research Lab, UC Berkeley
27 February 2025

What is DraCor (dracor.org)?



What is DraCor?

- Digital research infrastructure for European drama (and beyond) from the Antiquity to present
- Showcase for the “programmable corpora” concept
- **Core idea:** make dramatic texts machine-readable, comparable, and interoperable
- Open data, open source, community-driven development

23	9	4,000+	21
Published corpora	In progress	Plays	Languages

A growing corpora collection

Language-based corpora

FreDraCor
GerDraCor
EngDraCor
RusDraCor
DutchDraCor
ItaDraCor
HunDraCor
SweDraCor
PolDraCor
UDraCor
SpanDraCor
GreekDraCor
RoDraCor
SpanDraCor
RomDraCor
AlsDraCor
YiDraCor
TatDraCor
BashDraCor
NeoLatDraCor
CzeDraCor
EstDraCor
GeorgDraCor
HeDraCor
IndiEDraCor

Author-based corpora

CalDraCor
GerShDraCor
ShakeDraCor
IbsDraCor

National corpora

AmDraCor
ArDraCor

Precondition: TEI Standard

TEI = Text Encoding Initiative — the international XML standard for scholarly text encoding

```
  <sp who="#gretchen">
    <speaker>MARGARETE.</speaker>
    <lg>
      <l>Nun sag, wie hast du's mit der Religion?</l>
      <l>Du bist ein herzlich guter Mann,</l>
    </lg>
  </sp>
  <sp who="#faust">
    <speaker>FAUST.</speaker>
    <lg>
      <l>Laß das, mein Kind! Du fühlst, ich bin dir
  gut;</l>
      <l>Will niemand sein Gefühl rauben.</l>
    </lg>
  </sp>
```

Example: Goethe's Faust, pt. I (1808)

Why TEI?

- Structured markup of speakers, stage directions, acts, scenes
- Machine-readable dialogue, character data, metadata
- Enables consistent cross-corpus comparison

Emma Lazarus

Q240959

b. 1849, New York City

d. 1887, New York City

DEMO: https://dracor.org/am/lazarus-dance-to-death

The DraCor API

API = Application Programming Interface — standardised, structured access to all corpus data via URL queries

`…/corpora/ger/plays/goethe-faust…/spoken-text?gender=FEMALE`

All speech by female characters in Goethe's Faust

`…/corpora/shake/plays/hamlet/metrics`

Network metrics from Shakespeare's Hamlet (JSON)

`…/corpora/rus/plays/chekhov-vishnevyi-sad/stage-directions`

All stage directions from Chekhov's Cherry Orchard

`…/corpora/ita/metadata/csv`

Full metadata for the Italian corpus (CSV)

Full documentation: dracor.org/doc/api

Ecosystem

🌐 Web Interface

dracor.org — browse corpora, explore plays, download data, inspect networks interactively

🔗 Linked Open Data

Plays and authors linked to Wikidata, GND, VIAF — SPARQL-queryable knowledge graph + ontology

📡 API

Fully documented, versioned API (v1) for programmatic access to all corpus data + wrappers for R and Python + Claude MCP

🛠 Documentation

Tutorials (as Colab Notebooks), ODD guidelines, upcoming textbook

📦 Spin-offs

Experimental adaptation of the DraCor infrastructure for other domains: PoeCor, EcoCor

📁 Open Source

All corpus data and infrastructure code on GitHub: github.com/dracor-org

Community

DraCor Core Team (Vienna, 2024)

DraCor community at the first DraCor Summit (Berlin, 2025)

Summary

- a digital research infrastructure/ecosystem for the study of European drama
- reliable, expandable corpora in multiple languages+ highly functional interfaces
- open data, open source, open community
- ideally, an interface between analogue and digital philology

Website	GitHub	API Docs	Paper
dracor.org	*github.com/dracor-org*	*dracor.org/doc/api*	*doi:10.5281/zenodo.4284002*

DraCor for DH research

More than 70 publications have used DraCor data since 2015.

See dracor.org/doc/research

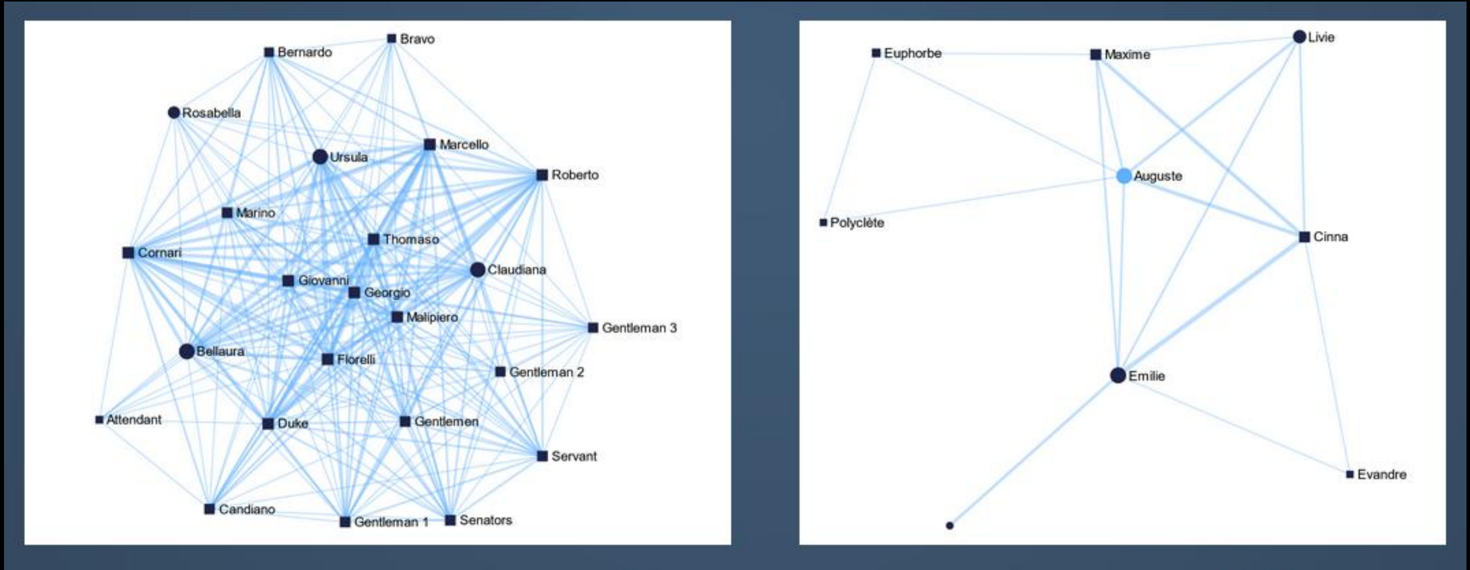
An example:

Evolutive Dynamics in Early Modern European Drama
(PhD project, Universities of Potsdam and Padua, 2021-2024)

Forking Paths: Evolution and Divergence in Early Modern European Drama (out in late 2026)

Research question

How did different European dramatic literatures develop their own distinct formal features during the early modern era?



Shirley, *The Gentleman of Venice* (1639)

Corneille, *Cinna* (1639)

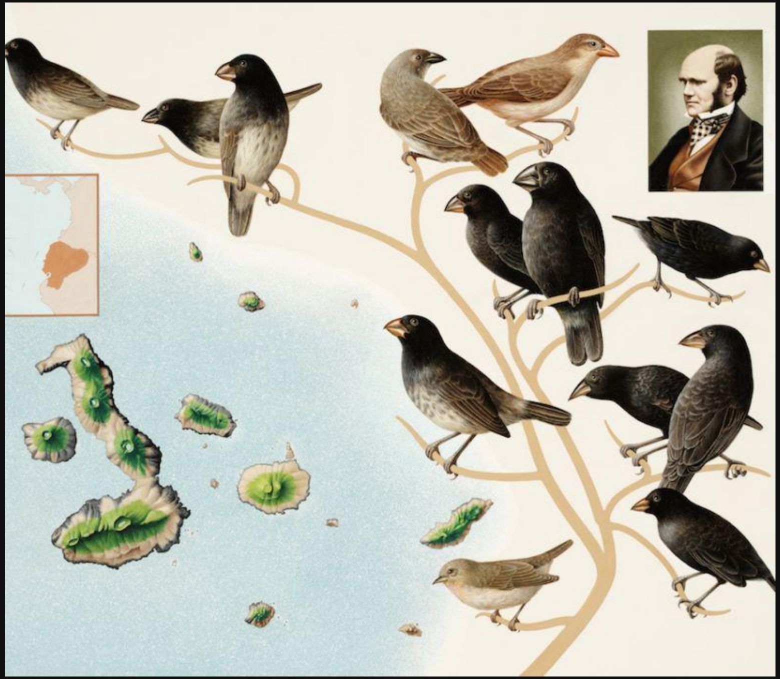
Theoretical framework

Theoretical Model (Moretti 1994)

- Literary evolution mirrors biological speciation
- European space as an 'archipelago' of national sub-spaces
- Each tradition has specialised in a distinct formal variation during the early modern era

Competing Explanations

Küpper (2018): Drama as a cultural net — plots, characters, motifs circulate across traditions
Clubb (1990): Diffusion of 'theatregrams' — emphasising cross-European unity over diversity
Drama aesthetics: ‘irregular’ drama (Spain, England) vs. ‘regular’ drama (Italy, France) (cf. Lotman's typology)



Corpus: EmDraCor (Early Modern Drama Corpus)

150
Plays

5
Languages

1561–1710
Time Span

Corpus Design Principles

- Purposefully non-canonical selection
- Balance between representativeness and practicability
- Leverages DraCor infrastructure via Dockerisation

Operationalising drama

Key Components of Drama

- Dialogue
- Characters
- Plot

Approach: Quantitative Formalism

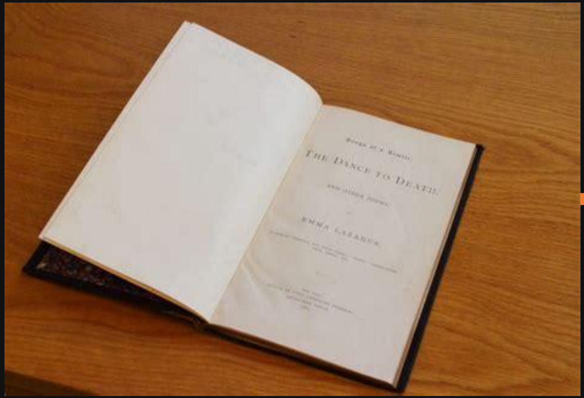
- Content- and language-agnostic
- Form-oriented structural analysis
- Metrics collected via DraCor API or self-computed following various studies

Category	Examples
Network	Size, density, avg degree, diameter, protagonism
Cast & Speech	Avg chars/scene, speech intensity, gendered speakers
Size	Acts, segments, word counts, verse/prose lines
Plot	All-in index, final scene size, drama change rate

Play Embeddings:

Metrics collected into multidimensional feature vectors capturing different structural dimensions holistically.

Operationalising drama



{41, 3, 14, 11, 3, 0, 0, 1, 3306, 16919, 16461, 473, 6.571428571428571, 0.505494505494505, 0.7699649556792411, 14, 11, 1, 46, 0.0494505494505494, 0.6608466660481829, 0.2431203018946239, 0.926829268292683, 0.5714285714285714, 0.6391666666666669, 3.0, 0.5974025974025974, 0.8461538461538461, 0.3571428571428571, 0.2857142857142857, 0.3571428571428571, 4.99, 0.3571428571428571, 0.2857142857142857, 0.3571428571428571, 0.0175417232479634, 0.5493558858726955}

Is this reduction acceptable?
Cf. Krämer's (2023)
„cultural technique of flattening"

Things one can do with play embeddings

#1 Measure

#2 Visualise

#3 Decompose

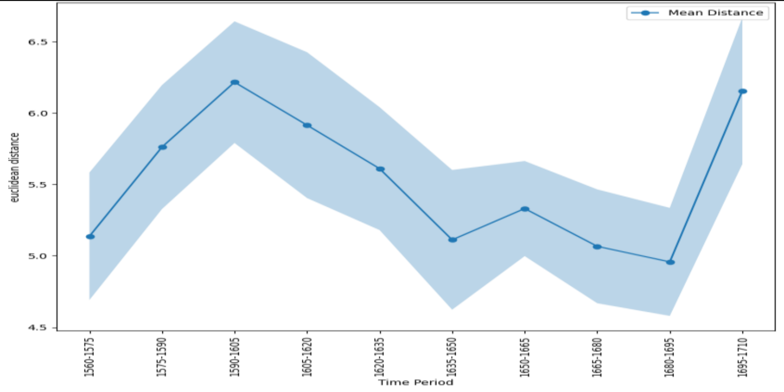
#1: Measuring vector distances

Ground Hypothesis

- Vector distances express degree of (dis)similarity between plays
- Progressive distancing of vectors → branching of dramatic traditions

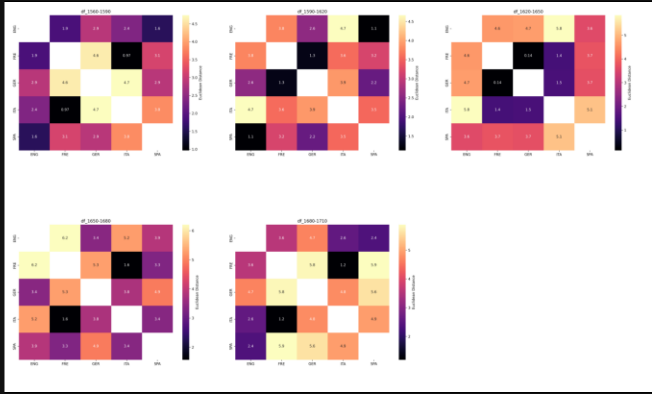
Method A — Pairwise Distances

- Euclidean distance computed per play pair
- Average distance = proxy for 'formal diversity'
- Result: convergence-divergence movement



Method B — Centroid-Based Distances

- National sub-corpora reduced to centroid vectors
- Cross-national distances plotted per time frame
- Result: partial evidence of growing divergence



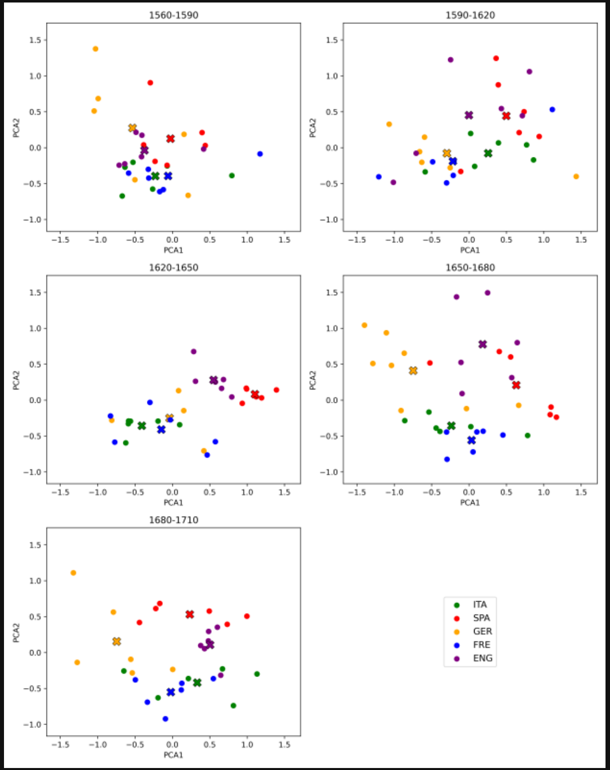
#2: Visualising vectors via PCA

Approach

- Dimensionality reduction (PCA) projects play vectors onto a 2D Cartesian plane
- Plays plotted per successive 30-year time frames (1560–1710)
- National traditions identified by colour; visual clustering assessed

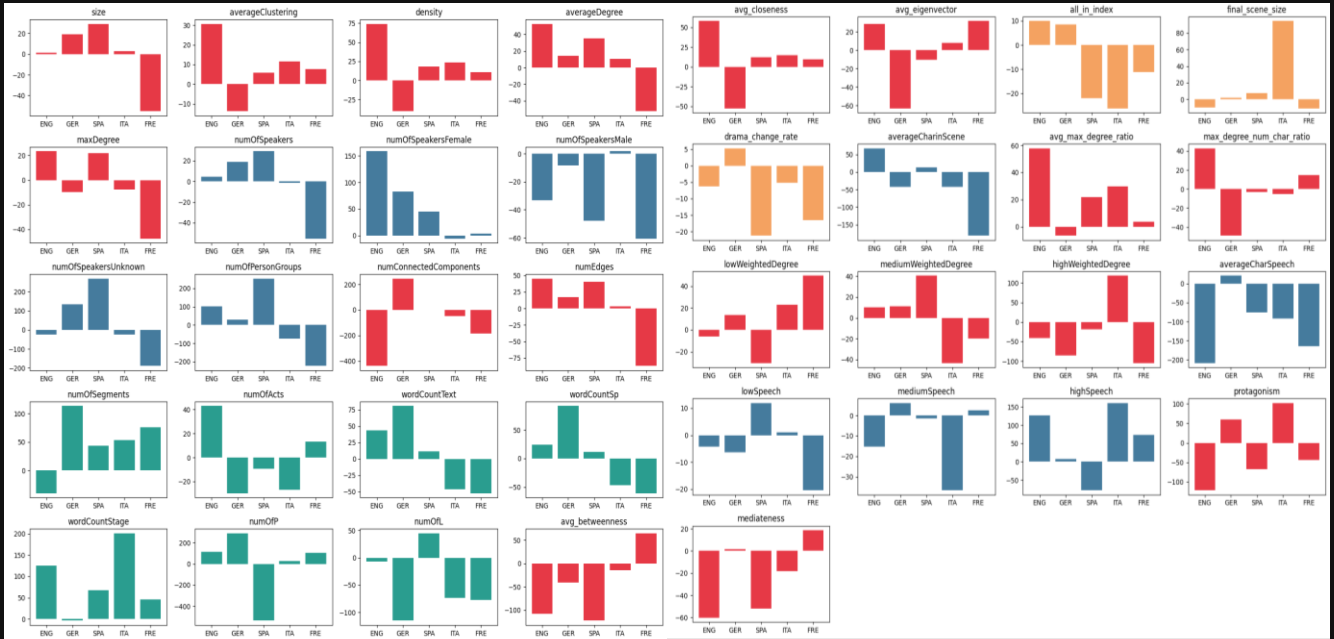
Findings

- Early time frames: plays highly intermixed — no clear national clusters
- Later time frame (1680–1710): emerging clusters



#3: Decomposing vectors

Approach: Track evolution of individual metrics across time per national subcorpus; compute absolute shifts; identify metrics highly distinctive of each tradition which help creating a quantitative profile of each dramatic culture.



Results:

ENG: Distributed networks, more actresses
GER: Sparser networks, more scenes
SPA: Even connectivity, collective voices
ITA: Protagonist concentration, more stage directions
FRE: Smaller casts, neoclassical economy

Limitations

On the poster: Boris Yarkho, a pioneer in the quantitative analysis of drama

Theoretical Framework

- Moretti's speciation model is a rigid, now outdated model of cultural evolution
- Other models (cf. Sobchuk 2023) might be more appropriate or complementary

Corpus Specifics

- Limited to 150 plays across 5 languages; selection bias
- Findings not completely validated against a reproduction experiment on larger French/English corpora

Operationalisation

- Metric set might not exhaustively represent 'drama'
- Formal features alone do not capture theatrical performance or reception, especially relevant in the early modern era

Results & Contributions

Key Findings

- Findings partially support Moretti's thesis of increasing formal diversification…
- …but suggest rather a 'horizontal' variation and branching than a straightforward 'vertical' evolution
- Metric distribution analysis provides strongest evidence of distinct national trajectories

Code and data:
https://github.com/lucagiovannini7/emdracor

Contributions

Corpus

Multilingual, open-access, TEI/XML-encoded corpus of 150 early modern plays

Theory

Empirical reassessment of Moretti's speciation model

Methodology

Play vectorisation based on formal features — a reusable pipeline for computational drama analysis

Thank you!

Some open questions for the discussion:

- How does form-based vectorisation of texts fare compared to language-based vectorisation?
- Is it really a viable approach for reconstructing literary history?

giovannini@uni-potsdam.de
lucagiovannini7.github.io