

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)?

The screenshot shows the DraCor website interface. At the top, there is a navigation menu with links for ABOUT, CORPORA, DOCUMENTATION, TOOLS, and RESEARCH. On the right, there are logos for GitHub and TEI 2022. The main heading is "DraCor - Open Infrastructure for Drama Analysis". Below this, there are five corpus cards, each representing a different language: French (Fre), German (Ger), English (Eng), Russian (Rus), and Calderón (Cal). Each card displays the following information:

- Corpus Name and Language
- Number of plays
- Number of characters (with gender breakdown: M, F)
- Number of text tokens
- Number of stage tokens
- Last update date and time
- User avatars and names associated with the update

At the bottom of the page, there are four columns of links: ABOUT, DOCUMENTATION, TOOLS, and RESEARCH.

Corpus	Plays	Characters (M/F)	Text Tokens	Stage Tokens	Last Update	User	
Fre DraCor (French Drama Corpus)	1,940	17,831 (M: 7958, F: 4247)	17,701,440	509,958 (16,295,137)	48,755 (393,275)	14.7.2025, 06:35:29	dc9c208
Ger DraCor (German Drama Corpus)	767	16,462 (M: 11632, F: 3430)	11,969,043	472,952 (11,407,483)	225,126 (1,387,768)	19.2.2026, 08:35:23	e948b7c
Eng DraCor (English Drama Corpus)	753	11,145 (M: 8564, F: 2080)	16,448,904	550,693 (15,057,801)	97,646 (447,851)	1.2.2026, 22:14:19	dbc4fc6
Rus DraCor (Russian Drama Corpus)	212	3,707 (M: 2608, F: 871)	2,316,995	119,333 (2,191,657)	49,440 (215,112)	1.2.2026, 16:39:25	f602fad
Cal DraCor (Calderón Drama Corpus)	205	3,406 (M: 1991, F: 1119)	2,838,830	121,432 (2,736,034)	23,933 (115,495)	1.2.2026, 02:54:39	60316cb

ABOUT
[What Is DraCor?](#)
[Corpus Registry](#)
[Credits](#)
[Get in Touch](#)

DOCUMENTATION
[API](#)
[API v0 \(legacy\)](#)
[Encoding Guidelines \(ODD\)](#)
[FAQs](#)

TOOLS
[pydracor](#)
[rdracor](#)
[SPARQL](#)
[ezlinavis](#)

RESEARCH
[Research Bibliography](#)
[Posters](#)
[Card Games](#)
[DraCor in Science Communication](#)

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

Published corpora

9

In progress

4,000+

Plays

21

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-vishnevyyi-sad/stage-directions
```

All stage directions from Chekhov's Cherry Orchard

```
.../corpora/ita/metadata/csv
```

Full metadata for the Italian corpus (CSV)

Ecosystem

Web Interface

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

API

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

Spin-offs

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

Linked Open Data

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

Documentation

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

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

dracor.org

GitHub

*github.com/dracor-
org*

API Docs

dracor.org/doc/api

Paper

*doi:10.5281/zenodo.4
284002*

DraCor for DH research

More than 70 publications have used DraCor data since 2015.

See dracor.org/doc/research

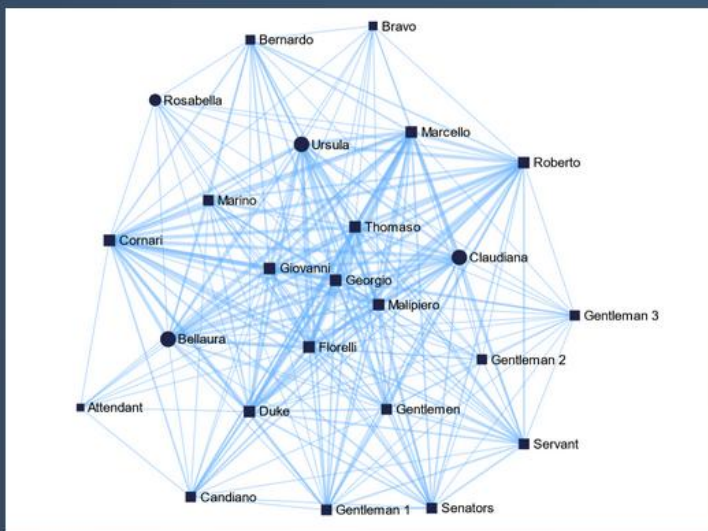
An example:

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

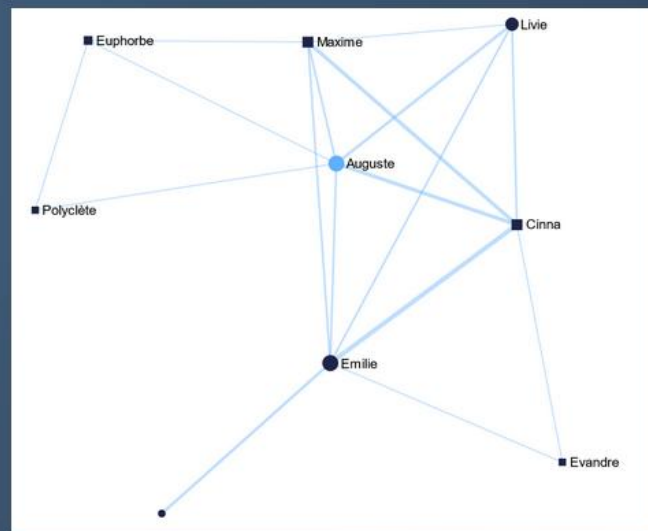
Forking Paths: Evolution and Divergence in Early Modern European Drama (out in late 2026)

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)

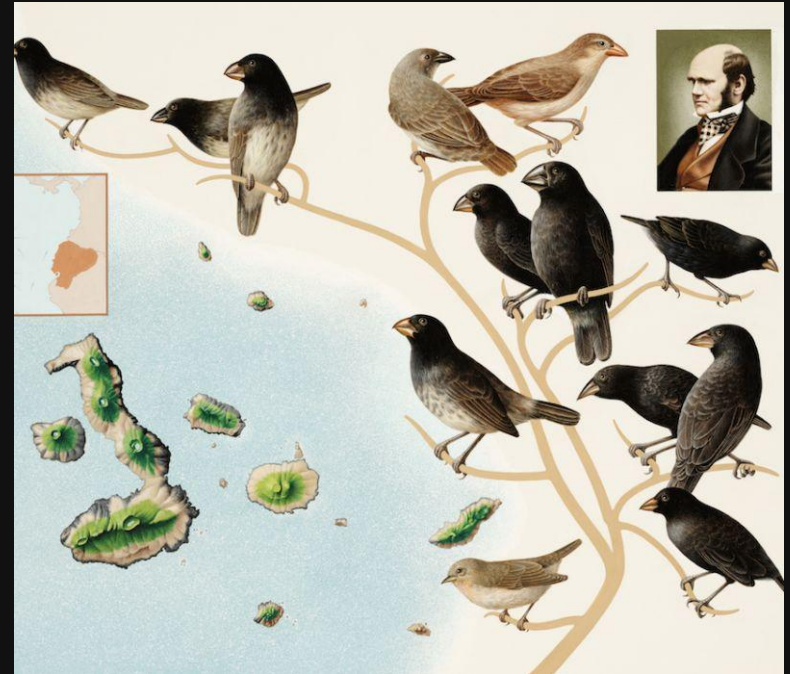
- Literary evolution mirrors biological speciation
- European space as an 'archipelago' of national sub-spaces
- Each tradition has specialised in a distinct formal variation during the early modern era

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

The screenshot shows a dashboard for the EmDraCor corpus. At the top, the title 'EmDraCor' is displayed in a dark blue box, with 'Early Modern Drama Corpus' written below it in a light blue bar. The main content area is white and contains several statistics, each with a large number on the left and a label on the right. The first statistic is '150 Number of plays'. The second is '3,148 Number of characters', with a sub-label '(M: 2241, F: 600)' and a 'person + personGrp' tag. The third is '3,012,955 Text tokens' with a 'text' tag. The fourth is '94,776 (Tokens)' with a 'sp' tag and '(2,767,411)' below the number. The fifth is '13,566 (Tokens)' with a 'stage' tag and '(84,025)' below the number. At the bottom, 'Last update' is shown as '19.2.2025, 16:32:44'.

EmDraCor	
Early Modern Drama Corpus	
150	Number of plays
3,148 (M: 2241, F: 600)	person + personGrp Number of characters
3,012,955	text Text tokens
94,776 (2,767,411)	sp (Tokens)
13,566 (84,025)	stage (Tokens)
Last update	19.2.2025, 16:32:44

Operationalising drama

Key Components of Drama

- **Dialogue**
- **Characters**
- **Plot**

Approach: Quantitative Formalism

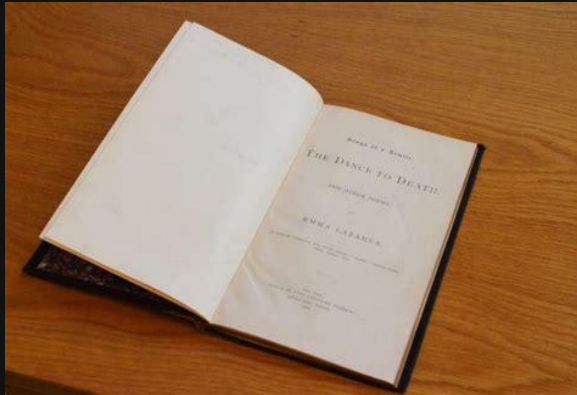
- Content- and language-agnostic
- Form-oriented structural analysis
- Metrics collected via DraCor API or self-computed following various studies

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

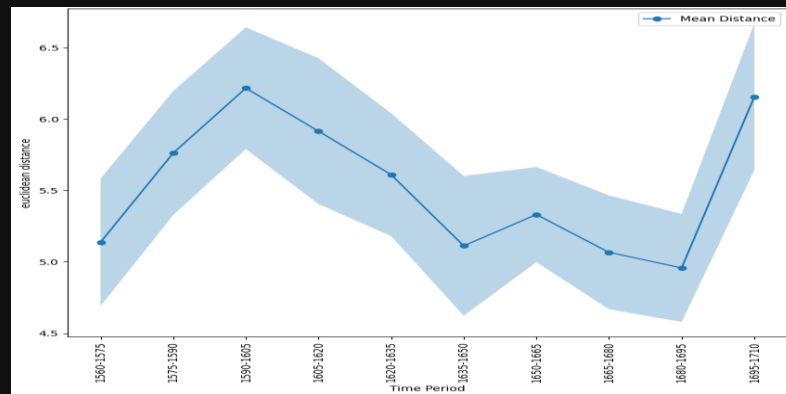
#1: Measuring vector distances

Ground Hypothesis

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

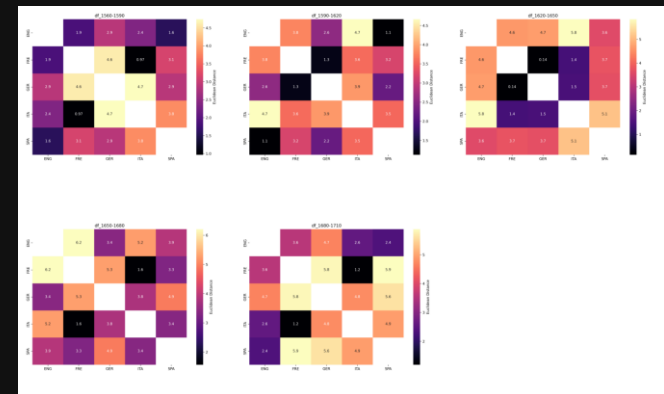
Method A — Pairwise Distances

- 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



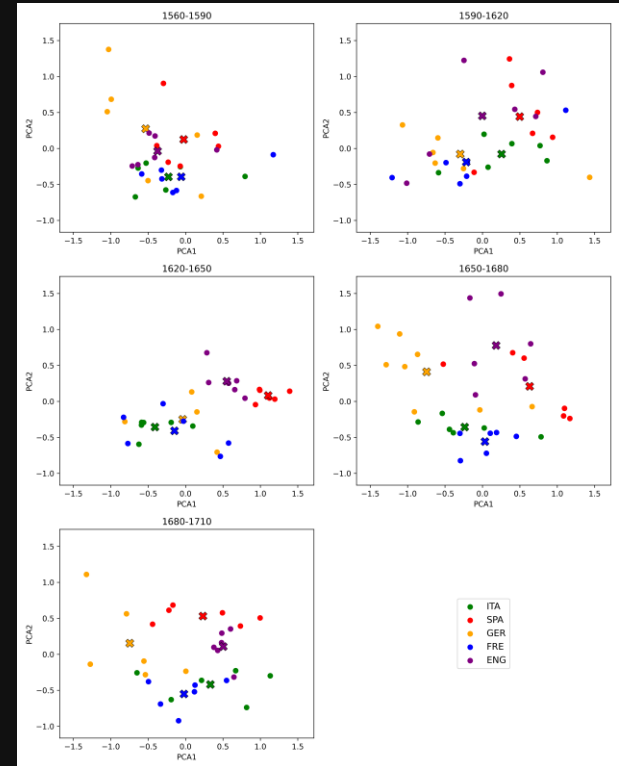
#2: Visualising vectors via PCA

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

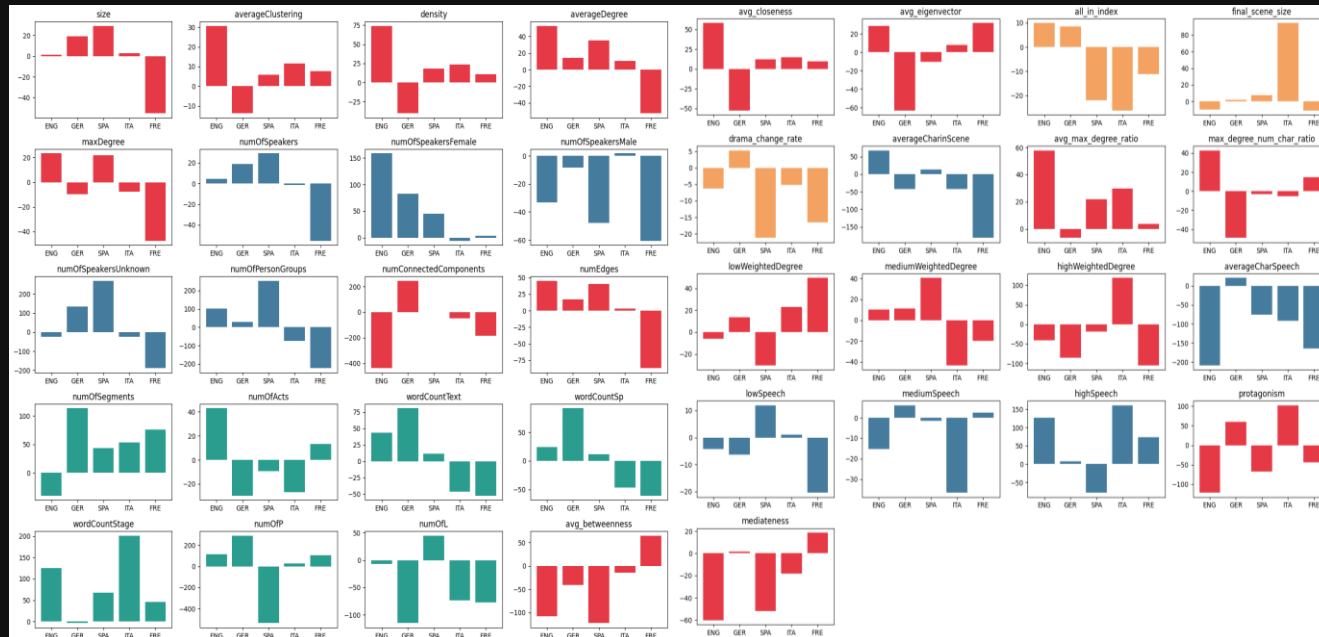
Findings

- 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](https://github.com/lucagiovannini7)