Methods

The pipelines described in the previous chapter produced a highly curated, machine-actionable corpus of early modern plays, whose standardised features now enable meaningful comparison across traditions. Before proceeding with the analysis, however, it is imperative to clarify which methodology is to be followed. Since, as Gius and Jacke (2022, 16) rightly underline, "computational approaches [should be] aligned with the theoretical background guiding the [original] analysis", I wanted to use an approach consistent with the framework originally employed in 'Modern European Literature'.

As evidenced in Chapter One, Moretti's focus there lay very clearly on the morphological variety of early modern drama – the study of "the component parts [...], their relationship to each other and to the whole", as Vladimir Propp (1968, xxv) would put it in the opening pages of his *Morphology of the Folktale*.[1] While still lacking a computational dimension, his approach foregrounded, to some extent, the paradigm of 'quantitative formalism' (Allison et al., 2011) in which he would spend the remainder of his career, and which he thus later summarised:

> Quantitative formalism: taking an aesthetic form and dismantling it down to its basic elements; transforming – or rather, let's say, reducing – a novel to its paragraphs, or a drama to a sequence of linguistic exchanges. Having done this, one is confronted with 'a second-level form', as Alex Woloch once wrote to me; a squared, abstract form, a

1 On the relation between morphology and literature see also Gilodi (2021) and Marx (2013).

> form of forms [...]. This is the eminently analytical vocation of quantitative formalism: to make everything rest on explicitly “local” interactions such as [...] the back-and-forth between one character and another. (Moretti, 2021, 385)

This book situates itself in the same tradition by treating plays as configurations of speech turns and scenic units rather than as linguistic content, i.e. by following a form-centred approach. Furthermore, such a strategy also makes it possible to circumvent two major features of the corpus, namely its multilingualism and historical character. EmDraCor contains texts in five different languages, spanning more than one century, with considerable linguistic variation within each tradition. Such inter- and intralingual differences would have represented a hurdle for directly applying natural language processing methods (see Piotrowski, 2012). To achieve reliable results in most tasks, it would often have required training multiple language-specific models from scratch rather than simply fine-tuning existing ones (cf. Manjavacas and Fonteyn, 2022), considering also that domain adaptation for literary corpora typically demands more than fine-tuning (Bamman, 2024).

In the following pages, I first briefly outline some possibilities for operationalising the concept of drama, recalling some older and newer approaches, and then present my own methodology for modelling EmDraCor plays.

Operationalising drama

From Moretti (2013b)’s seminal formulation – influenced, among others, by Percy Bridgman’s *The Logic of Modern Physics* (1929) and by Thomas Kuhn’s essay ‘The Function of Measurement in Modern Physical Science’ (1961) – to the most recent, DH-specific debate (see especially Krautter, 2022; Pichler and Reiter, 2022), operationalisation represents a core task in any quantitative approach to literature, and its role within every research pipeline cannot be overstated. “[B]uilding a bridge [...] from the concepts of literary theory, through some form of quantification, to literary texts" (Moretti, 2013b, 1), operationalisation is a delicate procedure which requires a balancing act between the specificity of the literary text and the algorithmic requirements of the machine.

In the context of computational drama analysis, perhaps the most important operationalisation practices concern the identification and quantification of some crucial dramatic features and the modelling of their relations. Defining what the key components of a play are, however, represents a complex theoretical problem in its own right, with its first discussion tracing back at least to Aristotle's comments on the six parts tragedy is made of.[2] One could say, with some approximation, that most scholarship has been working with what Kretz (2012, 105; cf. Kafitz, 1989, 32–40) called the three 'elementary components' of drama:

- *Characters*: the entities to be represented on stage, to be understood as deliberate literary constructs (cf. Pfister, 1988, 160–161, which for this reason prefers to use the word 'figure');
- *Plot*: 'the arrangement of the incidents' (*synthesis* or *systasis ton pragmaton*) happening during the play, which for Aristotle represents the most important component of drama;[3]
- *Dialogue*: the alternation of speech acts, which is often considered the first 'external' feature of drama one immediately recognises (Schlegel, 2018, 19) and thus a "primary distinctive feature" of the whole genre (Veltruský, 2017, 248).

The same tripartition can also be found in the Russian Formalist tradition, with Boris Yarkho – a major theoretical referent for this book, as the reader will shortly see – indicating 'dialogue' (*dialog*), 'characters' (*personazhi*), and 'action' (*dejstvie*) as the three distinctive elements of drama

2 "Necessarily then every tragedy has six constituent parts, and on these its quality depends. These are plot [*mythos*], character [*ethe*], diction [*lexis*], thought [*dianoia*], spectacle [*opsis*], and song [*melos*]" (*Poetics* 1450a, 5–10, trans. W.H. Fyfe).

3 "[…] most important of all is the structure of the incidents. For Tragedy is an imitation [*mimesis*], not of men, but of an action and of life" (*Poetics* 1450a, 15, trans. S.H. Butcher).

(Yarkho, 2006, 410). Of course, such narratology-oriented perspectives sideline once again the performative dimension of drama, which other researchers tried to recover,[4] but which also remains outside the scope of this work.

In any case, it can be argued that these three high-level components find a concrete expression in different textual aspects (content, style, structure…). Such aspects can in turn be investigated through a range of computational methods: as Willand et al. (2017) acknowledge, the most popular options in this sense have been the topic modelling of dramatic genres (e.g. Schöch, 2017b), the exploration of network structures (see *infra*), and the analysis of character speeches (e.g. through stylometry, see Šeļa et al., 2024; Savoy, 2023).

As mentioned earlier, my approach to operationalising the EmDraCor corpus is both content- and language-agnostic, and my methodology will be therefore grounded in form-oriented approaches. Before presenting my own strategy, I will briefly review previous similar approaches to drama analysis. Following and integrating the reconstruction by Fischer (2024), I will begin by outlining the exemplary contribution of three significant twentieth-century scholars – Boris Yarkho, Solomon Marcus, and Hartmut Ilsemann – whose pioneering work laid the ground for modern computational study of drama.

After briefly examining the current scholarly landscape, I will then devote a methodological excursus to literary network analysis, which has been central in past and present DraCor-based scholarship and will feature prominently in the experiments presented in Chapter Four. For this reason, it seemed helpful to briefly summarise its origins and discuss its strengths and weaknesses in modelling dramatic texts.

4 Bernhard Asmuth (1980, 4), for example, singles out the Aristotelian *mythos*, *lexis*, and *opsis* as the three most important formal elements of drama, while considering the characters (*ethe*) part of the plot itself. Accordingly, its own tripartite systematisation is based on the concepts of plot (*Handlung*), dialogue or character speech (*Figurenrede*), and 'sensory' performance (*sinnliche Darbietung*).

Stages in the development of computational drama analysis

Modern scholars agree in considering the formalist school, and more specifically its Muscovite incarnation,[5] a primary source of inspiration for quantitative formalist endeavours; the links between the two movements have been increasingly scrutinised in recent years (see e.g. Pilshchikov, 2020; Lvoff, 2021; Pilshchikov, 2022). This renewed interest within the digital humanities,[6] as Fischer et al. (2019a, 1) put it, is motivated by pragmatic reasons, since "we are now able to operationalise and automatise formalist research ideas, to reproduce them, to scale them up and to further develop methods along those lines". In other terms, modern methods seem now in a position of nurturing the seeds of 'computational thinking' which were already planted as early as the 1920s.

Moscow scholar Boris I. Yarkho (1889–1942), in particular, has been rediscovered as "a very early forerunner of contemporary research practice in the Digital Humanities" (Fischer et al., 2019a, 7). A multifaceted scholar well-versed in various languages and philologies, Yarkho made

5 As Pilshchikov (2020, *passim*) points out, the formalist movement has often been equated tout court with the St. Petersburg-based OPOJAZ (Society for the Study of Poetic Language) helmed by Shklovsky, Tynianov, and Eichenbaum, while its Moscow counterpart, i.e. the Moscow Linguistic Circle founded by Jakobson and frequented by Tomashevsky and Yarkho, has often been neglected. The difference between the two flavours of Formalism – which remained nonetheless intertwined thanks to the constant exchange of people and ideas – can be roughly summarised as follows: the OPOJAZ theorist advanced a more 'morphological' or 'functionalist' formalism, focused on texts (and thus to some extent 'qualitative'), while the Muscovites were rather involved with methodological issues, adopting a 'quantitative' perspective which is somewhat nearer to that adopted, in modern times, by the Stanford LitLab.

6 Two milestones in this sense can be recognised: the 2015 Stanford conference on 'Russian Formalism and the Digital Humanities' (program: https://digitalhumanities.stanford.edu/russian-formalism-digital-humanities) and the 2019 special issue of the *Journal of Literary Theory* on 'Moscow Formalism and Literary History', featuring new translations of works by Boris I. Yarkho, Mikhail L. Gasparov, and Maksim I. Shapir.

a significant contribution to drama analysis in two studies written between the 1920s and 1930s ("The Comedies and Tragedies of Corneille. A Study in the Theory of Genre" and "Speech Distribution in Five-Act Tragedies. A question of Neoclassicism and Romanticism"[7]), which showcased an original, quantitative-driven approach to the poetics of genre and of literary trends.

Another pioneer in the quantitative treatment of drama, from a different perspective, was Romanian mathematician Solomon Marcus (1925–2016). His approach, he claimed, went beyond the "simple statistics" done by previous authors, including the Russian Formalists, but aimed rather at constructing a full 'mathematical-linguistic' model of drama to which "various mathematical methods coming from game theory, graph theory, theory of sets, algebra of binary relations, probability and information theory, [and] formal grammars" could be applied (Marcus, 1999, 294–95).

Among his contributions, he notably developed the idea of a Boolean matrix to keep track of the changes in character configurations, based upon the virtual model of an 'unusual spectator' "who is only capable to observe the entrances and exits of the actors and monitor their co-occurrences on stage without listening to what they say" (Fischer and Skorinkin, 2021, 520). Marcus' whole approach to drama analysis was also notable in its strong functionalism: "the whole mechanism of a dramatic work is seen as a machine, with whose help I can understand better the functionality of a dramatic work" (Marcus, 1971, 150).

Marcus' configuration matrix was later commented upon in the work of leading German structuralist Manfred Pfister (Pfister, 1988, 171 ff.),

7 Both studies are appended to the modern Russian edition of Yarkho's magnum opus, the *Methodology for the Exact Study of Literature* (Yarkho, 2006). "Speech Distribution in Five-Act Tragedies" has also been translated into English (Yarkho, 2019) and its workflow has been computationally reproduced and tested by Wendell (2021). An English translation of "Corneille's Comedies and Tragedies" is still missing but its workflow has already been outlined and/or digitally reproduced by Sonkin (2022), Wendell (2021), and Šeļa (2020).

and later played a central role in the work of one of the first actual practitioners of digital drama studies, Hartmut Ilsemann, as exemplified in his paper 'Computerized Drama Analysis' (Ilsemann, 1995). There, he introduced an experimental but already quite complete set of programs for the quantitative study of plays, which translated ideas by Marcus, Pfister, and Wilhelm Fucks[8] into (computational) practice. In particular, software developed by Ilsemann and his collaborators[9] allowed for computing several statistics related to stage configurations and speech acts, and visualising them in several formats.

While more 'technical', Ilsemann's 1995 paper touched also upon some key points which remain debated in computational drama research to this day: the criticism against statistical methods, which seem to refute the uniqueness and irreproducibility of the work of art (12), the necessity of "a large data background" (13) to establish "a reliable typology of dramatic forms" (21), the belief that "quantitative statistical relations are capable of expressing qualitative aspects of dramatic plays" (14).

Inspired by structuralist and formalist thinkers such as Yarkho, Marcus, and others, and heralded by the path-breaking work of Ilsemann and colleagues, computational drama studies have become throughout the years an important component within the larger ecosystem of digital literary studies, both in quantitative[10] and in qualitative terms.

This vitality has also manifested in the establishment of several research projects dedicated to creating and annotating corpora and developing reusable programming tools. In the German milieu, for exam-

8 German physicist who worked on the application of empirical-quantitative methods to the humanities (see e.g. Fucks, 1968, 1971).

9 For a brief technical account of the inception of Ilsemann's original program see http://www.shak-stat.engsem.uni-hannover.de/progdoku.html. For its successor, the *Internet Drama Analysis Program* (IDAP), whose main purpose was providing statistical data on English Renaissance plays, see Ilsemann (2006).

10 See e.g. the increase in the share of drama-related papers in the *Index of Digital Humanities Conferences*, (Weingart et al., 2024).

ple, the long-standing DraCor initiative[11] has been over time joined by projects such as *Quantitative Drama Analytics* (*QuaDramA/Q:TRACK*, see Pagel et al., 2023[12]) and *Emotions in Drama* (*EmoDrama*, see e.g. Dennerlein et al., 2023[13]). Furthermore, specific fora have been established to allow scholarly exchange within the subfield, such as the ongoing series of *Computational Drama Analysis* workshops (Cologne 2022, Berlin 2025).[14]

Excursus: dramatic network analysis

Assuming the *Computational Drama Analysis* workshops as barometers for recent trends in computational drama research, one immediately notices how the methodology of (literary) network analysis is still playing a central role in current investigations. First popularised by Moretti himself in 'Network Theory, Plot Analysis' (now in Moretti 2013a, 211–40), and then more thoroughly developed among others by Trilcke (2013), the modelling of dramatic structures as networks remains indeed one of the most popular practices in computational literary studies, with a rich scientific production exploiting its potential for tasks such as investigating characters and plots, assessing literary theories, and more (see Labatut and Bost, 2020, 26–27; Trilcke et al., 2024, 2).

The remarkable success of this methodology in the subfield of drama studies – part of a wider 'network turn' (Ahnert et al., 2021) in the humanities – can be explained by pragmatic and conceptual factors. On the one

11 Emerging from the grassroots *DLINA* project (https://dlina.github.io), DraCor has later been supported by the pan-European *Infrastructure for Computational Literary Studies* (https://clsinfra.io).

12 https://quadrama.github.io/. See also the software (*DramaAnalysis*, *DramaNLP*) they developed for drama analysis: https://quadrama.github.io/tools.en.

13 The same research group also developed *Katharsis*, a "tool for drametrics" (cf. Romanska, 2014) which is based on Marcus' and Ilsemann's work and available at https://lauchblatt.github.io/Katharsis/sa_selection.html; see Schmidt et al., 2019.

14 The proceedings of the first workshop are collected in Andresen and Reiter (2024a); selected papers from the second one are forthcoming in the *Journal for Computational Literary Studies*.

hand, theatrical texts are well suited for computational network analysis, since they almost invariably present a rigid structure which can be easily transformed into a network where characters are represented as nodes and their relations as edges. Such relations could be modelled with different layers of precision, ranging from the creation of basic co-presence networks[15] to networks based on interactions, family relationships, or other complex connections (e.g. information flow, see e.g. Andresen et al., 2022).[16]

On the other, network representation offers a "particular kind of abstraction", which is "uniquely capable of working across scales" (thus supporting both micro- and macro-analyses) and at the same time "[i]ntuitive to grasp and yet mathematically complex" (Algee-Hewitt, 2017, 752). In this perspective, network properties are assumed to mirror some crucial features of the works – density may correlate with plot complexity, or centrality distribution may reflect character hierarchies. Accordingly, a comparison between different networks can yield insight into literary history and cultural development.

Furthermore, this methodology proves especially effective when deployed on multilingual corpora, because its "abstract modelling and the possibility of a (more or less) 'purely' topological comparison makes the comparative analysis highly independent of any domain- or language-specific origins of the networks" (Trilcke et al., 2024, 2).

Applications of network analysis to dramatic texts, however, raise methodological challenges as well. In an early phase, it was not uncommon to find papers which considered drama only as a test ground for verifying mathematical or statistical principles, and thus disregarded the works' literary specificity. These investigations, which often seemed to be little more than "a challenge for computer scientists" (*ibid.*, 7), often

15 This type of network – the easiest to extract automatically, since it only requires character identification and scene segmentation within the text – is the one natively featured in DraCor.

16 Such networks require further preprocessing of the source text (e.g., involving NLP pipelines to recognise speakers and addressees) or task-oriented annotation.

failed to offer significant contributions to computational literary criticism.

Similarly, one needs to problematise the role of social networks analysis (SNA), which initially acted as a bridge from (pure mathematical) network analysis to literary network analysis (LiNA). While providing an important inspiration, concepts and theories from SNA should not be applied slavishly to fictional texts: even though some properties of real-life networks can be found also in literary ones (Labatut and Bost, 2020, 47–48; see also Waumans et al., 2015), "[t]here is no reason to suppose that the writer tries to mimic actual social relationships when producing the work of fiction", and "numerous character networks extracted from fictions do not exhibit realistic topological properties" (Labatut and Bost, 2020, 5).

Accordingly, one should always assess the significance of network metrics within the context of the literary work one investigates, and not at some abstract level. Otherwise, as Peer Trilcke has warned,

> a literary network analysis which, when reflecting on its foundations, disregards the objection that it may miss the specific ontology and historicity of its object [will] hardly be able to deliver reliable textual empirical results from the viewpoint of literary theory. (Trilcke, 2013, 207)

In concrete terms, this also means that any attempt at literary network analysis should start from a reflected, thorough operationalisation of dramatic structures. It should, in other words, define clearly some key formal features of a play, namely, what constitutes an exit from or entrance to the stage, what constitutes a character, and what does it mean for two characters to have an interaction (Jokisch and Rojas Castro, 2025).

Modelling EmDraCor plays

In accordance with these methodological reflections, the first step in modelling dramatic texts would consist in identifying the features most suitable to describe them. Past scholarship has often focused on developing new measures for capturing what they consider key aspects of textual structures, but it is important to remember that creating such measurements should not be a goal in itself. Introducing novel metrics should rather provide an objective hermeneutical advantage over existing ones or conventional non-computational methods. Furthermore, it ought to be kept in mind that

> [m]easurements can be taken of any quantifiable aspect of a text, but figuring out the significance of that metric to an understanding of the text, or better, mapping that metric onto a preexisting critical concept (such as style, plot, or theme), is crucial to making sense of what is being measured. (Algee-Hewitt, 2017, 759)

Some examples from previous work might help frame the question. Yarkho's study of Corneille's plays, for example, was based upon 30 features describing various structural aspects, which he later cut to 15 measures he considered the most significant for his research goal, i.e. effectively differentiate between comedies and tragedies.[17] His subsequent paper on Classicism and Romanticism relied instead on only four features, all related to characters. Interestingly, Yarkho's quantitative-oriented, but still 'analogic' (i.e. pre-digital) approach led him to employ in his analysis eminently qualitative features as well. One of the parameters he employed in his essay on Corneille, for example, was 'saturation' – a blend of metrics indicating, among other things, "the percent of personages that act in their own interest" and "the ratio of the number of actions on stage to the number of actions that occur behind-the-scenes" (Gasparov, 2017, 141–42).

Other Russian (neo)formalists, reviewed in Wendell (2021, 31 ff.), followed Yarkho's example and developed their own features for the study of specific works. Revzina and Revzin (1971), for example, introduced metrics such as 'character connectedness', 'dramatic engagement rank', and 'probabilistic link', which anticipated several later DH endeavours;

17 The full list of features can be read in Wendell (2021, 39–40).

their approach was later updated by Sapogov (1974), whose proposal for calculating speech distribution is also implemented in the DraCor frontend (cf. Börner and Trilcke, 2023, 64 ff.). Sperantov (1998), for his part, focused on stage directions in Russian verse tragedy, which he explored through measures like frequency, average length, presence within verses, lexical diversity, etc.

Modern computational approaches have kept developing their own features as well. Work done by the DraCor group, for instance, has sought to tackle the progressive structuration of plot, following earlier attempts by Xanthos et al. (2016). In particular, Trilcke et al. (2017) devised several 'event-based' measures, which "are used to identify or characterise a particular point in time within a drama", and 'progression-based' ones, which "allow a general characterisation of the transformation of a dramatic network" (Fischer et al., 2017, 2–3). Thanks to these metrics, they were ultimately able to identify low- and high-dynamic plays within DraCor.

In his quantitative study of English early modern drama, Algee-Hewitt (2017) tried instead to build upon standard network measures of centrality, like eigenvector and betweenness, and developed two advanced metrics for the analysis of drama: 'protagonism' and 'mediatedness'. While the first one is meant to "measur[e] the tendency of the play to concentrate the function of the protagonist in a single character [...] or to distribute it among multiple characters", and to show the existence of an alleged "core/periphery divide" among the plays' characters (763), the second one embodies the "need for a character to bridge the factions and to mediate between the subnetworks" (767). Gauging the chronological development of these two metrics allows the author to draw some conclusions about literary evolution and genre history as well – e.g. tragedies and historical dramas becoming more morphologically similar to comedies (771).

While these last two examples are representative of the tendency to centre metric development on network-related aspects of drama,[18]

18 The reasons for this are thus spelled by Algee-Hewitt: "If the network, as a visualization, compresses the dynamic action of the play into a schematic morphol-

it could be argued that only a broader pool of features, able to cover as many aspects of a play as possible, might be truly useful for a deeper comprehension of literary history and cultural evolution dynamics. While an all-encompassing operationalisation of drama is virtually impossible, it is possible and advisable to employ multiple methods to capture different structural aspects of drama – not only network analysis, but also scores capturing size, speech distribution, plot development, etc.

In modelling the plays from EmDraCor, I thus propose to go beyond the comparison of a restricted number of measures and rather collect as many of them as possible in feature vectors, which I will then use as hermeneutical tools for examining intertextual relationships and literary history at large. Vectors, as is well-known, are mathematical entities that represent a quantity characterised by both magnitude and direction. In computer science, and by extension within the digital humanities, vectors are more generally defined as a set of numbers arranged in a specific order or orientation (e.g.: v = {1, 2, 3, 42}). The most commonly used vectors within computational literary studies are word embeddings, which aim to capture the meaning of words or text based on their distributional properties.[19]

In the context of computational drama analysis, one can follow a similar logic to build *play embeddings*, i.e. vectors containing a variety of measures embodying different textual aspects of the dramatic text. While not completely novel, this approach has in the past concerned only individual components of drama, such as characters. Solomon Marcus, for example, stored the appearances of characters as vectors

ogy in which all of the interactions (relationships, conflicts, romances, murders) are part of a static representation, then the metrics from the distributed character network reduce these multidimensional visualizations into statistics that represent, in minute detail, a single aspect of the morphology of the drama" (Algee-Hewitt, 2017, 756–57).

19 Following the famous precept by Firth: "You shall know a word by the company it keeps" (Firth, 1957, 179). For an introduction on applications of word embedding models in digital literary studies see Schöch (2022); some theoretical considerations in Dobson (2021)

and computed their 'difference' via Hamming distance (Marcus, 1970, 293; 300). More recently, there have been at least two notable attempts to computationally model characters in German drama:[20] Fischer et al. (2018) used count- and network-based character measures to describe quantitative dominance relations, while Krautter et al. (2023) employed a composite array of features[21] to investigate character types such as 'the virtuous daughter' or 'the tender father'.

To my knowledge, the only recent works which attempted formal vectorisation of *whole plays* are Wendell (2021), which however deployed it only where necessary to digitally replicate Yarkho's study on Corneille, and Szemes and Vida (2024), which applied it to solve a genre classification task. Their pioneering investigation, aimed at distinguishing comedies from tragedies within GerDraCor and ShakeDraCor, employed both standard network analysis metrics and some new features developed for the specific task.[22]

While the supervised classification procedure the authors performed did not find a "striking [structural] difference" (177) between the two genres, they could nonetheless single out some properties which were highly predictive of generic alignment – as speculated already by Ilsemann (1995, 16). Their results appeared thus to "confirm the existence of a 'genre fingerprint' that shapes the dramatic structure of tragedies and comedies" (Szemes and Vida, 2024, 175), and that playwrights may (un)consciously choose to adhere to or depart from. The algorithm was

20 For a similar, exploratory attempt on all DraCor corpora see Giovannini and Skorinkin (2025).

21 Their features cover nine different character-related aspects: text statistics, network measures, stage presence, semantic fields, topics, action verbs, sentiment scores, character information and drama information.

22 The metrics Szemes and Vida (2024, 171–175) employed were average clustering coefficient, density, average path length, diameter, maximum betweenness, average degree – maximum degree ratio, maximum degree – number of characters ratio, speech distribution (high/medium/low), weighted degree distribution (high/medium/low), average character speech, average character per scene, number of connected components. Some of these measures will be reused and explained here as well.

also rerun removing the plays' last act, which is generally treated as a strong genre predictor,[23] but the variation in results was limited.

In Giovannini and Skorinkin (2024a), we reprised Szemes and Vida's method and fine-tuned it in order to explore the features of opera libretti. In this study, we were not interested in a pure classification task, i.e. sorting out libretti from non-libretti, but we aimed at assessing whether libretti possessed some peculiar formal properties and how they evolved across the centuries. While results of this early experiment were not conclusive, confirming known challenges of modelling dramatic texts through structural features *only*, they also revealed some formal patterns which were typical of libretti, especially those with a non-comic background. Even more importantly, that paper represented a test ground for workflows and methodologies later employed here, especially in terms of feature selection protocols and visualisation workflows.

Before moving to describe the actual composition of my vectors, however, a brief reflection on the epistemic validity of this method is warranted. Compressing texts into a series of numerical values purported to represent some of their salient aspects, and then comparing these 'bags-of-numbers' to make assertions about literary history, could be easily called out as another example of the – at the same time – hubristic and reductionist mindset often attributed to digital humanities.

23 Comedies often end with a wedding where all the characters are present, thus making the network denser, while tragedies often end with the death of many characters, originating a less dense network. This observation is often traced back to Bernard Shaw, who did however complain about this simplistic view: "Now neither tragedies nor comedies can be produced according to a prescription which gives only the last moments of the last act. Shakespear [sic] did not make *Hamlet* out of its final butchery, nor *Twelfth Night* out of its final matrimony" (Shaw, 1921, xciv). The first evidence of this *topos* I could find is in Byron's *Don Juan*: "All tragedies are finish'd by death, / All comedies are ended by a marriage" (III, 9).

I would argue, however, that drama vectorisation[24] is just another expression of the 'cultural technique of flattening' German philosopher Sybille Krämer assumes as foundational for DH endeavours (Krämer, 2023a; see also Krämer, 2023b), and which is actually shared both by 'standard' humanistic inquiry and modern computational approaches. According to her, the opposition between traditional modes of scholarly interpretation, represented by 'deep' hermeneutics, and data-driven, 'shallow' DH endeavours is indeed misconstrued, given that humanities have always been using 'superficial' representation of cultural items as their research objects:

> Historically, modes of working in the Humanities emerged in close liaison with operations involving artificial flatness. What is represented on a surface always forms an empirical, and therefore countable, fact. Thus, the number has always played an essential role in the generation and representation of knowledge in the Humanities: Concordances, library signatures, catalogs of works, historical timelines and dating tables, and diagrammatic inscriptions of all kinds bear witness to this. (Krämer, 2023a, 15–16)

Consequently, computational methods differ from traditional methods only in the extension of their 'flattening' procedure, i.e. in the way in which they transform cultural objects into machine-actionable data. Following the model of the natural sciences, however, they also require thorough documentation of modelling choices as a standard practice, and this attitude helps the critical re-evaluation of assumptions that often remain hidden in non-digital hermeneutics.[25]

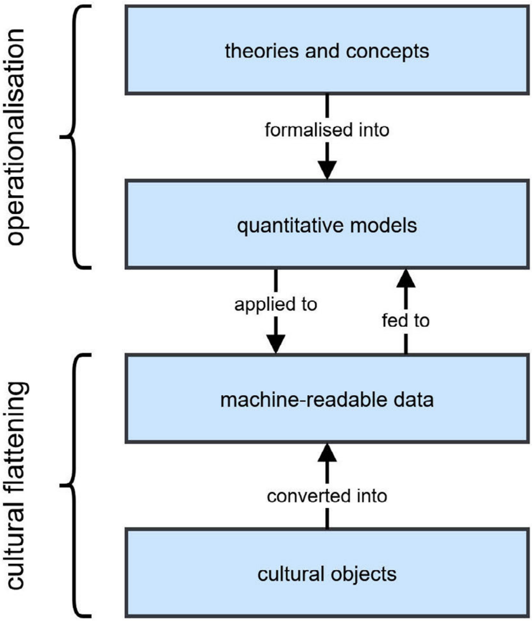
---

24 I use here the term in its broader sense, referring to "the process by which objects are represented as vectors" (Dobson, 2021, 82). This should not be confused with its more common meaning in machine learning, where it refers to accelerating code execution by performing computations on a set of values simultaneously rather than iterating through individual elements.

25 "The Digital Humanities can be understood as an extrapolation and radicalization of the productive dimensions of the 'cultural technique of flattening'. Its critical role towards the traditional Humanities may consist in making explicit

On an abstract level, one could even argue that the cultural flattening described by Krämer is a necessary methodological counterpart to another core DH practice, operationalisation: as tentatively schematised in Figure 4, these two practices converge in providing computational scholars respectively with the data and the interpretative models they need.

Figure 4: A minimal flowchart outlining two interlaced DH practices: cultural flattening and operationalisation.



just that kind of 'surface orientation' which – mostly concealed by the 'depth rhetoric' of hermeneutics – has always formed a genuine dimension in scholarly work (Kirschenbaum, *The Remaking of Reading*), but which has mostly remained a blind spot in Humanities' self-understanding" (Krämer, 2023a, 8–9).

Within this perspective, drama vectorisation should be interpreted not as a reductionist attempt to overtly simplify the complexity of theatrical texts, but rather as a continuation of formalist morphological thinking with modern computational methods. The 'flattening' of the plays' structural properties into vectors thus opens up new possibilities of inquiry both in the exploratory and confirmatory research paradigms (Gius and Jacke, 2022), without removing the need for 'scalable' readings to make sense of the patterns the vectors show.

A typology of features

The feature vectors I built to represent EmDraCor plays include several components from different sources. Most of them derive from the DraCor API,[26] whose endpoints allow one to access 96 features distributed across four levels: corpus, play, segment, and character. At the play level, it is possible to access 43 unique features[27] containing metadata and structural information (cf. Börner and Trilcke, 2023, 37–41). Out of these features, about half are non-numeric or not related to the text in itself (e.g. composition/performance dates), and thus were excluded from my vectors.

Additionally, I enriched the play embeddings with some further centrality measures (computed on the basis of data from the API) and with metrics developed in the context of three significant CLS studies I already mentioned (Algee-Hewitt, 2017, Szemes and Vida, 2024, Trilcke et al., 2017). I ended up with 40 features; four additional, non-numeric descriptors (slug,[28] language, genre, and year as string) were retained for

26 https://dracor.org/doc/api.

27 Those are the features one can see for each play in the corpus-wide metadata table downloadable for each corpus. The API ontology actually names 59 play-level features, but some of them are excluded from the metadata table, e.g. due to redundancy (such as the English translations of play title and author name).

28 A slug is the last part of a clean URL, akin to a filename. In DraCor, it constitutes a human-readable, unique identifier for a play and usually consists of the author's surname and the shortened title (e.g., *marlowe-doctor-faustus*).

identification purposes but excluded from the actual computations performed with the vectors.

Relying directly on the DraCor API to obtain most features was convenient but also entailed accepting its algorithmic choices, which are not always immediately understandable. A case in point is the computation of the 'average path length' and 'diameter' metrics.[29] When a graph has multiple independent components (i.e. is disconnected), the functions used by the metrics service return averages over within-component paths only. The resulting values are computable but semantically ambiguous: an average path length of 2.3 for a three-component graph is not comparable to the same value for a fully connected one. Given this uncertainty, I removed these two features to avoid interpretive confusion, bringing the final tally to 38.

The initial choice of so many features was motivated by an admittedly inclusive approach, in the belief that collecting a higher number of features would help give the vector more 'depth', i.e. more informational power. Such a strategy closely mirrored that adopted by Krautter et al. (2023, §39): on one side, the "features [chosen] cover as broad a field as possible, in order to exploratively determine which information is actually relevant" for the research question, and thus "complementary features can be found, each of which contributes new information that is not covered by other features". On the other side, the features picked are not novel, and their usage is grounded in previous theoretical and applied research, as illustrated in the following pages.

Nonetheless, this implementation deviates from standard machine learning procedures in dispensing with the feature selection phase, i.e. not removing features which are correlated (such as density and average number of characters per scene). This can be motivated by two reasons. On one hand, as in Szemes and Vida (2024, 173), most correlations between pairs of drama features do not necessarily mean the relationship between them is exclusive, i.e. "they cannot be said to describe the same phenomenon or phenomena [...] in a deterministic relationship". On the other hand, some techniques I will later employ natively deal with

29 Cf. also https://dlina.github.io/Subgraphs/.

feature redundancy and multicollinearity, while others focus on the independent interpretation of individual features, where multicollinearity becomes less relevant. Accordingly, I chose to maintain such a "multidimensional approach when it comes to the quantitative analysis" of plays (Fischer and Skorinkin, 2021, 528).

Furthermore, given the rich tradition of previous computational research on drama, I do not believe, as Fischer et al. (2017) do, that one must develop additional measures in order to enhance computational understanding of drama. Rather, I advocate for a rational and well-informed reuse of metrics already established, and consider developing new metrics only if the informational power of the existing ones proves insufficient. Nonetheless, I concur with Fischer and colleagues in considering "decisive to hold interpretations of [any] measur[e] against the backdrop of historical drama poetics" (4) – i.e. to employ metrics which are meaningful for one's specific research object (in this case, early modern drama).

The EmDraCor vectors contain a diverse selection of features, which can be roughly grouped into four thematic areas: network data, cast features and speech distribution data, size-related data, and data linked to plot dynamics. Altogether, these metrics contribute to an understanding of dramatic form, insofar as they concretely operationalise aspects of the three 'elementary components' of drama previously introduced. Network metrics and cast features, for example, address character relations and configurations; speech distribution data quantifies the structure of dialogue; plot dynamics approximate aspects of *mythos*.

As already mentioned, most metrics were obtained directly from the DraCor metrics service, while some supplementary measures are also calculated on data from the API itself. Furthermore, a significant amount of work was dedicated to re-computing metrics from third-party studies, which in most cases meant "developing reverse-engineered code that would functionally approximate the analysis that the author[s] must have performed" (Schöch, 2023, 392, italics mine). Indeed, the original scripts were either written in other programming languages or not available at all,[30] and thus some approximation could not be avoided.[31]

Detailed 'model cards' explaining each feature's source, meaning, and computation process can be found in Appendix C. For the reader's convenience, a list of the vector features is also provided in Table 1.

Beyond classifying them according to their category (network, cast, size or plot), vector metrics can also be systematised according to other categories. More specifically, they can be arranged on two axes – one which goes from *static* to *dynamic* features, and another one that goes from *global* to *local* features. Not all features, of course, fit neatly on these axes; this taxonomy is proposed here as an interpretive aid rather than a rigid schema.

First, a distinction can be introduced between metrics which immediately express a property of the whole play (e.g. its number of female characters), and are thus static, and metrics which attempt instead to capture the work's temporal dimension, usually by reflecting developments throughout the plot. While in principle all features can be sampled repeatedly to reflect, even if coarsely, variation across time (e.g. by computing the average number of any measurable component per scene), some metrics have been designed with this specific purpose in mind. A notable example in this sense is represented by the already mentioned

30 Plot measures developed in Trilcke et al. (2017) are not directly retrievable through API calls since they are dynamically generated via JavaScript to be displayed in the 'Speech distribution' tab for each play. Szemes and Vida's (2024) code is written in R and available on *GitHub*: https://github.com/SzemesBotond/drama_cluster_genre. Algee-Hewitt's (2017) script was written in R but is not publicly deposited.

31 For example, in a few plays the computation of some metrics failed, mainly due to API call issues resulting in null values. Since the play embeddings needed to have the same dimensions, I had to fill these missing values. Although there is no optimal solution for this, I adopted a common approach in data science by using the mean value for that metric within the play's own subcorpus. To make an example: if the metric 'mediumSpeech' could not be computed for Baron's *Mirza*, I filled it with the average 'mediumSpeech' value from the English subcorpus. All these changes are thoroughly documented in the code used for processing the data.

studies by Trilcke et al. (2017) and Fischer et al. (2017), which exploited network theory measures in order to investigate plot development, modelling the "[t]he structure of a literary text [...] as a changing sequence of network states" (Fischer et al., 2017, 1).

Table 1: Vector components, grouped by type.

Type	**Features**
Network	Size; Average clustering coefficient; Density; Average degree; Maximum degree; Average betweenness; Average closeness; Average eigenvector centrality; Number of edges; Number of connected components; Ratio, average degree to maximum degree; Ratio, maximum degree to number of characters; Weighted degree distribution (high/medium/low); Protagonism; Mediatedness
Cast and speech	Average characters per scene; Average length of character speech; Speech intensity distribution (high/medium/low); Number of speakers; Number of speakers by gender (male/female/unknown); Number of collective speakers
Size	Number of acts; Number of segments; Word count, whole text; Word count, spoken text; Word count, stage directions; Number of prose lines; Number of verse lines
Plot	All-in index; Final scene size; Drama change rate

Secondly, metrics can be classified based on the levels of scale they encompass. While some of them describe a 'local' component in the play, such as a character or a scene, others directly relate to the play as a whole. Network metrics, for instance, include both types: centrality measures mostly profile individuals in their relationship with the rest of the cast, while clustering measures, by contrast, offer a broader picture by describing all-encompassing features like group cohesion, information dispersal, etc.

Ultimately, the taxonomies outlined in the previous paragraphs underscore the heterogeneity of the metrics collected. This is a key feature in the design of the EmDraCor embeddings and finds its inspiration, once again, in the work of Yarkho – who argued that quantifying literary form requires "deriv[ing] consolidated (mean) indicators for all features under analysis, regardless of their heterogeneity (as it is impossible to prove, for example, that the percentage of dialogues is a feature more noticeable or, alternatively, less noticeable than the number of scenes) " (Gasparov, 2017, 144). Accordingly, the vectors I created aim to capture dramatic form as broadly as possible, combining metrics of different types and scales precisely because no feature alone can serve as a sufficient proxy for a whole play.