Corpus

The issue of corpus is crucial for any investigation of problems of literary history which aims at moving beyond the canon-centred paradigm and deploying 'distant' (Moretti, 2013a) or 'scalable' (Mueller, 2020) readings. Accordingly, scholars in computational literary studies have devoted considerable energy to the identification and description of 'good practices' one should follow to assemble corpora for digital investigation (cf. Desagulier, 2017; Mrugalski et al., 2022; Schöch, 2017a). Key criteria to be fulfilled in literary corpus building, as outlined by Herrmann and Lauer (2018, 136 ff.), include representativeness, balance, and fidelity, but one should also reflect upon the theoretical framework of the research, the possible inclusion of translations, and the need to adhere to FAIR principles (Wilkinson et al., 2016).

Furthermore, researchers need to consider another factor, which plays an even more decisive role when working on early modern materials: text availability. Mechanisms of cultural transmission are notably complex, and it is often difficult to reliably estimate how many works – in my case, how many plays from the early modern age – have survived the 'slaughterhouse of literature' (Moretti, 2000), and how many have been subsequently digitised, encoded, and annotated to be fully machine-actionable.

An example might help illustrate this point. In 1755, a group of Venetian intellectuals, coordinated by librettist Apostolo Zeno, completed the revision of the *Drammaturgia*, an authoritative catalogue of Italian drama first compiled two centuries earlier by a Roman erudite, Leone Allacci. Thanks to the recent conversion of the *Drammaturgia* into a

database (Giovannini and Gallucci, 2024; see also Gallucci and Giovannini, 2025),[1] it is now possible to use this resource to roughly gauge the extension of early modern Italian theatre.

The catalogue lists about 6,000 plays from 1449 to 1751, but out of its more than 2,000 authors, fewer than one third have a corresponding item in *Wikidata*, suggesting that most writers and titles are likely ignored by modern scholarship. While some texts might be at least digitised in large digital archives such as *Google Books* or *Internet Archive*, only a tiny fraction of them is encoded in TEI-XML and available in the standard digital library for Italian literature, *Biblioteca Italiana*,[2] which was used as the source for the Italian Drama Corpus (ItaDraCor).

Given this often-unsatisfying data landscape – partially shared by all early modern European literatures, though in highly varying proportions – assembling a corpus suitable for exploring the evolution of early modern drama presented multiple challenges. On a general level, the corpus building strategy was tailored to Herrmann and Lauer's (2018) advice (based on Mueller, 2020), according to whom the corpus building process needs to be oriented first and foremost by the research question one wants to address: "the size of the corpus on which a study is based and the degree of abstraction, the research design and the research methodology can only be determined on the basis of the respective research question" (138).

Therefore, the following pages report on all steps leading to the creation of the so-called *Early Modern Drama Corpus* (EmDraCor), including the procedures for text retrieval and the pipelines for corpus building and deployment. I will start, however, by stating some general principles behind the selection of the actual texts, pertaining to their chronological and linguistic collocation and to their coverage.

1 The *Allacci Digitale* database can be accessed at https://allacci-digitale.github.io/index-en.html.

2 The search engine (http://www.bibliotecaitaliana.it/catalogo) returns as few as 162 plays between the fifteenth and eighteenth century.

Corpus selection

Time frame

As famously pointed out by Lotman, "[i]t is virtually impossible to divide periods according to dates", since "human culture is a dynamic system", and "[a]ttempts to locate stages of cultural development within strict temporal boundaries contradict that dynamism" (Bassnett, 2002, 48–49). As a consequence, literary periodisation can be indeed seen as "the product of a process of generalization and synthesis, within a historical and social horizon, of the small-scale critical and interpretive practices that characterize the study of literary text" (Ciotti, 2022), whose validity is however confined to the hermeneutic framework one assumes.[3]

In his essay, Moretti spoke explicitly of 'Baroque' tragedy, employing a somewhat vague label which originated in art history before finding its way into music studies and eventually literary criticism.[4] As it often happens with categories borrowed from other disciplines, the notion of Baroque is far from uncontested, experiencing relevant semantic shifts in different countries, and appears unsuitable as a catch-all term for literary movements as different as the French *classicisme*, the Spanish *Siglo de Oro*, or the English theatre from its Elizabethan inception to its later offshoots.

Therefore, I prefer in this work to employ the broader notion of 'early modern drama'. Within the usual time span it is often associated with (roughly from mid sixteenth century to the end of the seventeenth), the choice of the year 1561 as the corpus' start date was admittedly informed

3 On this topic see also Wellek and Warren (1949, ch. 19) and Meneghelli (2013, 3).

4 The history of the concept of 'Baroque' – first appearing in the writings of Jakob Burckhardt and later popularised by Heinrich Wölfflin in his *Renaissance und Barock* (1888) – cannot be fully outlined here. For its specific significance in literary criticism, see at least Hatzfeld (1955), Wellek (1946).

by symbolic reasons. Indeed, during the Christmas and New Year festivities of 1561–62, the Inner Temple in London hosted the inaugural staging of *The Tragedie of Gorboduc*, a collaboration by Thomas Norton and Thomas Sackville, which is often considered both the first printed English tragedy and the first tragedy in blank verse (Winston, 2019, 185).

Inaugurating the 'golden age' of the Elizabethan theatre, this play could be assumed as exemplary for the broader, Europe-wide movement towards autonomous dramatic forms with some degree of independence from classical models. In this sense, *Gorboduc* assumes a paradigmatic role, explicitly acknowledged in Moretti's first version of the essay (Moretti, 1993, §3), and its staging year seems therefore fit as the lower chronological boundary for this investigation.

The choice of the cut-off date for the corpus proved challenging as well. On one side, Moretti's essay does not stop at early modern drama, and after examining the 'polycentrism' of Baroque tragedy it moves forward to investigate the birth of the 'Republic of Letters' concept and the spread of French literary hegemony and neoclassicism. On the other side, the 'branching' process is described as quite swift, but no precise time frame for it is established:

> The most significant transformations do not occur because a form has a lot of time at its disposal: but because at the right moment – which is as a rule very short – it has a lot of space. […] [Form] achieve[s] rather quickly a stable structure, which remains unchanged for decades, until it becomes sterile and disappears". (Moretti, 2013a, 13)

To shed clarity on this, a *longue durée* analysis of Moretti's claims would require including texts covering not only the decades where the branching happened, but also the following ones, where national form consolidated. Accordingly, a 150-year arc (i.e. from 1561 until 1710), was considered long enough to capture effective variations in dramatic form while remaining within the boundaries of the common definition of 'early modern' age. Furthermore, the endpoint signals, if not the end of the differentiation process, at least a first relevant caesura, marked

by such a major shift as the triumph of French neoclassicism all across Europe.

Scope

While Moretti's narrative of the evolution of European early modern drama moves seamlessly between authors and works from different cultures, supporting his argument with examples from Calderón, Racine, and Shakespeare, his focus always remains on the "three great western nation states" (i.e. Spain, France, and England) and "the German and Italian territories" (11).

Now, reducing the whole of European literature to some main 'centres of prestige' is not uncommon in comparative studies, but nevertheless problematic. At least in the case of early modern written culture, such simplification might find some justification in raw statistics such as the size of book markets. According to data provided by Buringh and Van Zanden (2009, 419), for example, between 85% and 93% of the books printed between 1550 and 1700 stemmed from one of the five linguistic areas explicitly mentioned by Moretti.[5] While partial and not specifically drama-oriented, such data offer a rough proxy for cultural infrastructure and readership, and thus support focusing on such a selection of linguistic areas.

From another perspective, however, such choice might still be questionable if one aims at constructing a comprehensive narrative of European literary history. For example, it would be fair to object to the exclusion of the Netherlands, both from a quantitative – book production from 1600 onwards there begins to dwarf the Spanish one,[6] – and qualitative perspective (ignoring the likes of Vondel, Hooft, Vos, or the *Nil Volentibus Arduum* society).[7] The contributions to early modern drama made

5 Buringh and Van Zanden's figures are derived from the *Global Historical Bibliometrics* database (https://www.iisg.nl/bibliometrics).

6 On the momentous growth of the Dutch printing industry in the seventeenth century see Hoftijzer (2001).

7 One could also argue that excluding Dutch drama from a comparative analysis of European drama may bias results toward Moretti's differentiation thesis, as

by other 'peripheral' language spaces, such as the Portuguese or the Polish one, seem less prominent, but could have also helped in getting a clearer picture of the whole.

Nonetheless, pragmatic considerations and stricter adherence to Moretti's framework, again following Herrmann's and Lauer's advice, suggested limiting this investigation to texts from the Spanish, German, English, Italian, and French traditions. At the same time, extending the investigation to further literatures remains a desideratum and could provide more layers of depth to any general argument about early modern drama.

Extension

Having defined the time frame and the linguistic scope of the corpus, the last step required determining its size. Such an operation translates into a balancing act between 'representativeness' and 'practicality' (Reppen, 2022, 32), i.e. between the desire to comprehensively cover a phenomenon and time- and resource-related constraints. These last issues are particularly relevant outside of the context of larger projects, such as EMOTHE – *Classics of Early Modern European Theatre* (Oleza and Tronch, 2020) or DraCor itself, which rely on distributed teams which collaboratively create and manage their corpora.[8] Furthermore, comparability of collections (Schöch et al., 2021, 15) has to be taken into account, since the five subcorpora I aimed to build needed to be homogeneous enough to allow productive cross-readings.

As anticipated, the question of availability of early modern works in digital formats plays a decisive role in defining the corpus' size. Despite massive advances in digitisation, many minor early modern texts are still

---

Dutch theatre is notable for its heavy international influences (e.g. from the Spanish *comedia nueva*, see Vergeer, 2026). On the role of the Netherlands as an exchange hub for dramatic models see also Bloemendal and Smith (2016, 17–21).

8 On the concept of 'distributed corpus building' in DraCor see Giovannini et al. (2023).

not available either in a standard scholarly format for critical editions, such as TEI-XML, or in simpler formats such as HTML or TXT. Furthermore, their distribution displays a noticeable variety across literatures, following obvious patterns of cultural prestige: for instance, properly encoded seventeenth century plays in Italian, beyond a handful of canonical authors, are rare, while there is an overabundance of digital resources on Shakespeare (Estill, 2019).

Given all these elements, multiple strategies for corpus building were taken into consideration; following the advice by Biber (1993, 243), I first focused on "a thorough definition of the target population and of the method of sampling" itself, before deciding the sample size. A first approach, for example, could have involved collecting wide lists of early modern plays from old and new bibliographic resources[9] and then performing chronologically stratified sampling on the titles to obtain the desired number of titles. This option, while having the advantage of countering canonical bias, was excessively time-consuming: drawing a truly random sample would have returned texts whose digital availability, even as scanned items without text layers, was far from assured.

Another option would have been relying on literary histories to identify works which, being already recognised as relevant and thus partially 'canonised', were likely to be already available in digitised formats.[10] While easier, such an approach would have had a negative impact on representativeness; if one follows a truly quantitative line of thought, any narrative of the evolution of European drama cannot be built only on a sequence of exemplary works, but rather requires engaging *at least to some extent* with what Cohen (2002, 23) famously termed 'the great unread'.

9 E.g., the *Cambridge Bibliography of English Literature*, vols. 1–2, for English.

10 See, e.g., *Canon 60*, a collection of Siglo de Oro Spanish plays edited by Joan Oleza and the EMOTHE team for the *Biblioteca Virtual Miguel de Cervantes* (https://www.cervantesvirtual.com/obra/canon-60-la-coleccion-esencial-del-tc12-teatro-clasico-espanol).

A third method, perhaps the most pragmatic one, consisted in exploiting existing digitised resources to their full extent and reducing new encoding tasks as much as possible – in other words, assembling what Mrugalski et al. (2022, 21) call "an opportunistic corpus". This translated into retrieving as many digitised plays as possible and then filling the gaps by manually encoding further texts from scratch. The negative implications of this strategy are clear, especially as it necessarily reintroduces some canonical bias. At the same time, as Annelie Ädel acknowledges, "it is […] very difficult not to include some element of opportunism in corpus design, as we do not have boundless resources", and therefore such an approach can be accepted where time- and resources-related constraints are present, and provided that "criteria for selecting material for the corpus be clear, consistent and transparent" (2020, 6).

Ultimately, the corpus size was set at 150 plays – a manageable size that ensures temporal coverage without exceeding practical constraints. While such an extension situates the corpus in an intermediate position – too big for an accurate close reading, too small for data-driven distant reading or cultural analytics (Manovich, 2020) – it nonetheless meets the requirements for a pilot study on evolutive trends within early modern drama.

Corpus retrieval

Once the preliminary corpus parameters were set, I moved to identifying and retrieving the actual texts. To do so, I first checked the relevant DraCor collections together with a number of databases offering open access to TEI-XML-encoded early modern plays, including the *Deutsches*

Textarchiv,[11] *TextGrid*,[12] EMOTHE,[13] *EarlyPrint*,[14] the *Oxford Text Archive*,[15] and *Théâtre Classique*.[16]

Figure 1: The number of plays written between 1561 and 1710 and available as TEI-XMLs in selected datasets.[17]

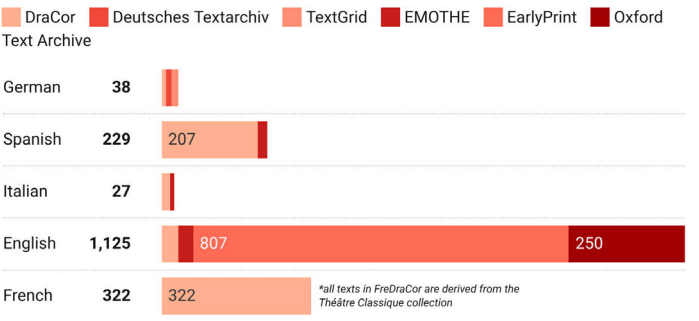


Figure 1 conveys immediately the imbalances in text availability across different collections:while there is plenty of TEI-XML plays in French and English, cutting the need for other sources almost to zero, the other language corpora are less resourced. Therefore, the next step required retrieving texts which were already available online but in formats other than TEI-XML (e.g. HTML, TXT, DOC).

11 https://deutschestextarchiv.de.

12 https://TextGridrep.de.

13 https://emothe.uv.es/biblioteca.

14 https://texts.earlyprint.org/works.

15 https://ota.bodleian.ox.ac.uk.

16 http://www.theatre-classique.fr. This was cross-checked with the *Dramacode* index (http://dramacode.github.io), compiled by Frédéric Glorieux.

17 These statistics refer to the time at which this research project started, i.e. April 2022. Since most of these corpora are 'living', the number of early modern texts has meanwhile increased: as of January 2026, for example, GerDraCor hosts 17 plays written between 1561 and 1710.

Most of these plays came from scholarly projects (such as the *Biblioteca Virtual Miguel de Cervantes*[18] or individual philological editions), public-oriented archives (like *Wikisource*[19]), or commercial databases (such as the ProQuest/Chadwyck-Healey's *Teatro Español del Siglo de Oro* collection[20]). Only as a last resort did I turn to performing OCR, correcting, and encoding scanned texts provided by platforms such as *Google Books*[21] or the *Internet Archive*.[22]

As far as the actual choice of titles was concerned, the main concern was ensuring that the whole 150-year-long time frame under investigation was evenly populated, with each period adequately represented. In practical terms, this translated into employing random stratified sampling to select ten titles for each decade (two for each language) from the available pool of texts. In the decades where no texts were immediately available, I used bibliographic sources (listed in Appendix A) to identify suitable titles to retrieve and encode from online archives; in this phase, some degree of hand-picking while filling these gaps was unavoidable.

Some additional criteria were enforced during the selection process. Where possible, I tried to avoid including anonymous works, and I checked the authors' distribution so that there were no more than three works for an individual writer. I also excluded operas and other dramatic texts which feature extensive singing or music (e.g. the *Singspiele* of the late German Baroque, the Venetian *drammi musicali*, etc.), preferring to focus on more established dramatic genres.

Finally, some issues common to all corpora might warrant some words. On one side, I was often faced with a lack of complete and/or reliable bibliographical information on some less-known works, as it often happens in the field of early modern studies. For example, there

18 https://www.cervantesvirtual.com/. The *Teatro clásico español* (TCE) collection from the BVMC contains 4388 items between primary texts and critical literature, available sometimes as HTML and sometimes as PDF (scanned images).

19 https://wikisource.org.

20 This database, accessible at https://www.proquest.com/teso, contains more than 800 plays by 16 different dramatists.

21 https://books.google.com.

22 https://archive.org.

might be difficulties in dating with certainty some works, such as plays published in collections only years after their composition or representation, and this would negatively impact any attempt at reconstructing an accurate genealogy of theatrical evolution. To address this issue, I consulted critical literature to collect the years of composition, first representation and publication for each play, which are used to determine the 'normalised' or reference year DraCor employs.[23]

On the other side, the issue of copyright needed to be taken into account. While most plays are in the public domain, some are derived from scholarly editions and thus still protected by copyright (Margoni and Perry, 2011). Since EmDraCor texts were from the start intended to end up in the main DraCor corpora – released under the *Creative Commons* 0 (CC0) license – I re-used scholarly editions only when their copyright had expired, or employed them only as reference while independently conducting transcriptions from public domain sources.

Corpus building

Encoding pipelines

Figure 2 schematically summarises the various procedures I developed for converting texts with different layers of markup (from well-formatted TEI-XMLs to plain texts from OCRs) to the standard DraCor format as described in its ODD encoding scheme.[24]

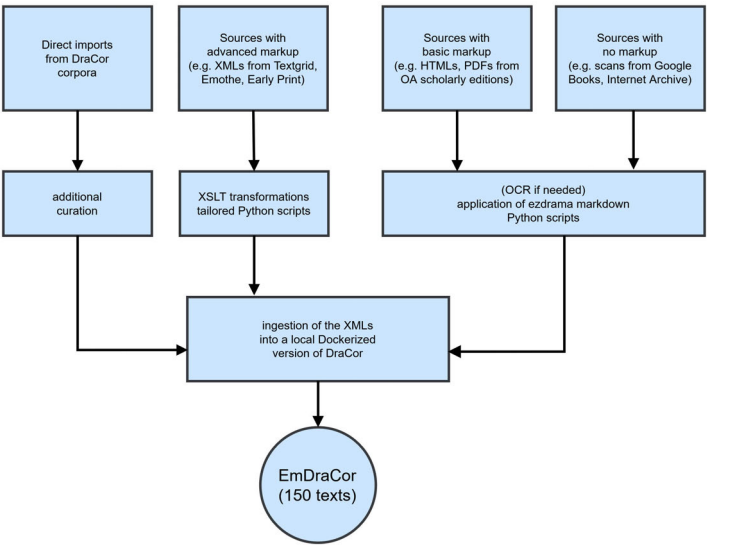
To begin with, extensive bibliographical research made it possible to locate more than one hundred texts already in TEI-XML format. Among them, about 40 were already featured in some DraCor corpus, and thus were only downloaded and subjected to some additional curation to ensure homogeneity. Other files, originating from various scientific projects on early modern drama, needed instead a complete refactoring of their markup to make them DraCor-ready.

23 See https://dracor.org/doc/faq.

24 See https://dracor.org/doc/odd.

Generally speaking, the transformation procedure consisted in two main tasks: the removal of excessive layers of semantic and/or graphical annotation (e.g. lemmatisation, part-of-speech tagging, facsimile layouts etc.), which are superfluous to my purposes, and the attribution – if not already assigned – of a speaker tag to each speech instance, which serves as a prerequisite both for literary network analysis and speech distribution statistics.

Figure 2: Corpus onboarding pipeline.



Tailored 'DraCorisation' pipelines were developed for converting the TEIs obtained from archives which provided a sizeable number of texts to the final corpus, such as *TextGrid*, EMOTHE, and *EarlyPrint*. The *EarlyPrint* corpus, in particular, presented a larger opportunity, as I collaborated with other DraCor team members not only to transform the thirty

plays needed for this project, but also onboarded the whole collection to create the *English Drama Corpus* (EngDraCor).[25]

The main challenge there lay in speaker disambiguation, i.e. identifying which character speaks each line. This problem proved difficult to take on, as early modern printing conventions were inconsistent, with speakers identified variously by name, role, abbreviation, or description, sometimes shifting within a single text. Resolving these ambiguities was essential for any network analysis or speech distribution metric, since the DraCor algorithms are based on the speaker tags each character is marked with. Once speaker attribution was stabilised through a dedicated tool,[26] the texts underwent XSLT transformation to produce DraCor-conformant TEI files. The EmDraCor texts were subsequently forked from the main project for independent curation, though they remain compatible with the larger EngDraCor architecture.

About one-third of EmDraCor plays, however, were derived from files in formats other than TEI-XML. Some of them still contained some semantic markup, such as the HTML files from *Wikisource* or the PDFs from various digital scholarly editions. Other texts, instead, were preserved only as scans in the digitised collections of *Google Books* and the *Internet Archive*, which allow downloading PDFs and plain text outputs of optical character recognition (OCR). While these transcripts were still messy, and thus required extensive post-correction, tests with other OCR tools such as *OCR4all* (Reul et al., 2019) did not show dramatic improvements in text quality; accordingly, I decided to work directly with the raw outputs from the aforementioned platforms.

For these lower-quality sources, I adopted a modified version of the conversion utility *EasyDrama* (Skorinkin, 2024). The principle behind its functioning is simple: rather than wrestling with inconsistent markup, one strips the text to plain TXT and annotates structural elements (such as act and scene divisions, speaker attributions, and stage directions) using a lightweight, Markdown-like syntax. The script then transforms this annotated plaintext into valid TEI-XML. This approach trades granular

25 https://dracor.org/eng; on its inception see Börner et al. (2023a).

26 https://dracor-org.github.io/epdracor-whois/.

markup for speed: while some basic markup information from the original HTML or PDF is lost, the conversion process becomes fast enough to handle dozens of texts. Given these advantages, I decided to process all non-XML sources through this pipeline, regardless of their original format.[27]

Editorial practice

The general goal of the DraCor project is not to produce flawless philological or critical editions, but rather to make dramatic works both readily available to readers and suitable for computational analysis, i.e. machine-actionable. Accordingly, editorial interventions on texts are most often cut to a minimum and confined to markup adjustments. In the case of EmDraCor, however, its comparatively small size made possible some minor fixes and improvements. Such editorial interventions chiefly regarded low-quality plain texts derived from OCRs, but some of them, like the improvement of segmentation, were applied to all texts.

On one side, even though the methods later applied to the corpus are language-independent, improving text quality still remained necessary to ensure the correct computation of certain metrics. Accordingly, I applied an extensive clean-up pipeline to all OCRed text. This started with the selection of the most recent copyright-free edition (often from the 19th century) as a source, in order to reduce OCR noise as much as possible. Since I did not have philological interests and I wanted to avoid remaining embroiled in the intricacies of digital scholarly editing,[28] I also settled for a normalised/modernised transcription as unobtrusive as possible, using (when available) the *Wikisource* style conventions as guidelines.

Among a few other interventions, I restored correct punctuation when lost during the OCR, normalised speaker names to ease their

27 Some novel approaches, not yet developed at the time of the corpus preparation, include semi-automatic, LLM-powered TEI encoding (Giovannini and Skorinkin, 2024b) and direct OCR-to-TEI pipelines (Dennerlein et al., 2025).

28 For an overview see Pierazzo (2020).

subsequent disambiguation,[29] and corrected some notable errors in markup (e.g. songs clearly sung by a character not being attributed to him/her, but rather encoded in a separate element, etc.).

On the other side, since most calculations performed by the DraCor algorithms are based on textual segmentation, the operation of dividing a text into units appeared crucial as well. Together with speaker disambiguation, it has indeed been often singled out as a major issue when trying to operationalise dramatic structures.[30] More generally, scene segmentation represents a long-standing issue both in natural language processing and computational literary studies, with the concept of 'scene' blurring into adjacent ones such as 'plot', 'topic', and 'event' (see Zehe et al., 2021, 3167–69).

Drama knows essentially two types of segmentation: the high-level 'act' division and the low-level 'scene' division. The usage of the former type, in particular, has fluctuated according to the poetics of each age, with the criteria for marking act breaks shifting over time: the *Oxford Encyclopedia of Theatre and Performance* speaks in this respect of "some combination of (a) an interruption of chronological continuity, (b) a change of setting, (c) a significant change in thought, mood or plot development" (Vince, 2003).

As far as literary theory is concerned, the genealogy of the concept of act starts with Aristotle's comments about the five stage units of Greek drama, upon which Horace later grounded his peremptory recommendation: "neve minor neu sit quinto productior actu fabula"[31] (Ruggerio, 1973, 173–74). Horatian influence, mediated through later grammarian Aelius Donatus, eventually came to dominate classicist-oriented theatre in the early modern age, but was not readily accepted everywhere.

---

29 Early modern texts often feature several different spellings and abbreviations for the same character. Thus, *Richard* might appear as *Rich.*, *Ri.*, *R.*, etc.

30 See e.g. the comments by Brainerd and Neufeldt (1974, 35 ff.) on the work by Salomon Marcus discussed in the next chapter.

31 "the play should not be either longer or shorter than five acts" (Hor. *Ars Poetica* 189–190, trans. T. S. Dorsch/P. Murray).

Here again, it is possible to see a split along the fault line previously introduced: French and Italian theatre, following the classical models mediated by the humanists, made use of acts, while other dramatic traditions followed the rather relaxed segmentation from medieval theatre, where pauses within the action were mostly linked to stage reasons (Lochert, 2018a, 191).

It is not surprising, then, that competing four- and three-act structures, based on different readings or willing departures from previous theorists, became available especially (but not only) in the English and Spanish milieus. Three-act structure, in particular, became the norm in Spanish drama following the precepts set by Lope in his *Arte nuevo de hacer comedias* (1609).[32] Even no segmentation at all remained an option: the majority of English printed plays between 1565 and 1594 had apparently no act or scene division (Howard-Hill, 1990, 142), and the segmentations later introduced in Shakespeare's plays were little more than artificial "editorial encrustation[s]" (Hirsh, 1981, 12). German theatre, for its part, was in the first phase quite irregular as well, with extra acts being freely added after the fifth when the text's length required it (Jordan, 1939, 399).

Given this heterogeneity of practices, it is fortunate that the standard segmentation span used by DraCor algorithms[33] is not the act, but rather the scene, a perhaps more 'natural' partition born out of the necessity to deal with changes in the characters' configuration on stage or to move around theatrical props. The number of characters which needed to be changed in order to identify a scene break, however, was again dependent on local dramatic conventions, as Lochert (2018b, 201–4) shows.

French neoclassical theatre, for example, often saw a character remaining on stage and being the *trait d'union* between different configurations or scenes, in a practice called *liaison des scènes*,[34] while the sub-

32 See vv. 211–212: "El sujeto elegido […] en tres actos de tiempo le reparta" ("Distribute the chosen subject in three acts of time") (Vega, 1914).

33 Technically speaking, they use the lowest-order (i.e. most nested) element, which most often corresponds to a with <type="scene">.

34 On this see Lochert (2018b, 204 ff.) and especially the full-length study by Douguet (2023).

stitution of a full set of characters with another marked an act change. Shakespeare, on the other side, tied scene sequence not to characters, but to changes in location (Ranke, 2010, 710, qtd. in Krautter, 2023, 273). These two concurring options, effectively tying scene division to character groups or to places, were variously received in other dramatic traditions.[35]

Most texts in EmDraCor already featured first- and/or second-level segmentation, usually from their original source. Works digitised and transcribed from historical editions, however, did not always have this markup in their printed editions, often for the historical reasons outlined above. Maintaining part of the corpus without a proper segmentation, however, would have hindered any meaningful comparison through computational tools, and thus required some fix.

Previous work in scene segmentation offered some inspiration. For example, Gius et al. (2019), using categories from Genette (1972), proposed this broad narratological definition:

> A scene is a segment of the *discours* (presentation) of a narrative which presents a part of the *histoire* (connected events in the narrated world) such that (1) time is equal in *discours* and *histoire*, (2) place stays the same, (3) it centers around a particular action, and (4) the character constellation is equal. (Zehe et al., 2021, 3169)

Drama scholars introduced other segmentation modes specifically tailored to theatre practice. For instance, Marcus (1970) introduced the concept of dramatic 'configuration', defining it as "the time frame during which no entry or departure from the stage takes place" (291). Jansen (1968, 77) proposed instead a new unit, the 'situation', whose boundaries

35 This crucial distinction was already noted by critic Heinrich Wilhelm von Gerstenberg in his *Briefen über Merkwürdigkeiten der Literatur* (1767): "It should be noted in advance that the English divide the scenes in a different way than the French. The former indicate a change in the stage, but the latter only indicate a change in the characters, for which the English have no special made-up word" (qtd. in Jordan, 1939, 403).

were defined by the change of characters and/or scenic elements; in doing so, he was again reprising, to some extent, the notion of scene as in French classical theatre.

Drawing on this, I opted for a simpler, pragmatic solution. Close reading of the unsegmented plays suggested that single entrances or exits rarely signal major dramatic shifts, while the movement of two or more characters typically marks a genuine change in stage configuration. Accordingly, a full scene change was added in the markup when both the following criteria were satisfied:

- At least two speakers left or entered the stage
- There is a stage direction unequivocally signalling such action (often shown by verbs of motion such as *salen*, *vanse*, *exeunt*, *escono*, etc.)

Such general principles, which try to establish a common baseline for texts from very different dramatic cultures, remain of course highly subjective. As Bartsch et al. (2023, §56) underline, "segmentation is by no means a purely formal matter but must always be thought of as a hermeneutical practice" which has a noticeable influence on text modelling.[36] Whichever solution is chosen, however, describing and motivating the criteria employed for segmentation has positive epistemological effects, favouring correct interpretation of the results and later re-evaluation of the methodology.[37]

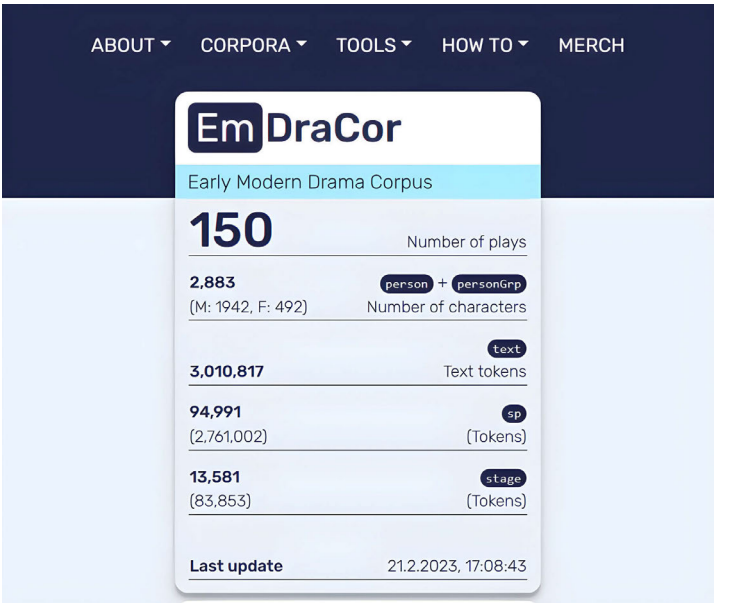
36 Bartsch and colleagues divide the segmentation process in three phases: "the cutting of a segment from the text continuum (zoning), the assignment of the segment to an analytical category (subsumption) and the determination of the depth of mental access on the basis of more or less contextual information (interpretation)" (§32).

37 "The digital access forces the disclosure of segmentation decisions, so to speak, and strongly promotes the definition and explication of criteria for these decisions. The algorithmic formulation as well as its implementation as a clear and discrete segmentation decision thus sharpens my concepts" (Bartsch et al., 2023, §74).

Deployment

Once all TEI-XML files were clean and properly encoded, I loaded them into a dockerised version of DraCor, following the procedure described by Börner et al. (2023b). This workflow uses the Docker technology to recreate a fully functional, locally hosted DraCor environment, complete with its API, frontend, and metrics service. This container can then be populated with TEI-XML plays from any sources, exported as an image, and later reactivated as needed.

Figure 3: Model card for the Early Modern Drama Corpus.



This procedure contributes significantly to addressing the 'reproducibility problem' often raised within the digital humanities (O'Sullivan, 2019; see also Chapter Four, 'Coda'), insofar as it freezes the corpus in a stable version on which one could later re-run the same original

experiments run here. Accordingly, the final product of this pipeline, EmDraCor (Figure 3), can be analysed using the full range of DraCor tools.