Appendix A

This appendix synthetically documents the sources, processing pipelines, and editorial decisions behind each of the five EmDraCor subcorpora. It complements the discussion of corpus design in Chapter Two by providing granular information on the provenance of the texts, the strategies used to fill temporal gaps, any exclusion criteria applied, and known issues that may affect downstream analysis.

French subcorpus (EmDraCor-Fre)

Sources	24 plays from FreDraCor (originally from *Théâtre Classique*, cf. Fièvre, 2007); 2 plays from *Wikisource*; 4 plays encoded from scans (*Google Books*, *Internet Archive*, *Gallica*).
Data landscape	Wide availability of texts thanks to FreDraCor (containing 552 plays from 1561 to 1710); some gaps before 1600.
Issues	FreDraCor's sheer size has delayed fine-grained curation and fixing of markup inconsistencies, mostly regarding wrong speaker attribution.
Gap-filling strategy	Missing texts retrieved through the *Répertoire des pièces de la Renaissance française, 1537–1615* (Bourassa, 2009) and the *CÉSAR (Calendrier électronique des spectacles sous l'Ancien Régime et la Révolution)* database launched by Barry Russell (cf. Bannister, 2009).

Processing	Plays from FreDraCor (XML): checked and corrected for speaker attribution errors, missing/wrong sex attributes, dating issues. Remaining plays (TXT, sometimes from OCR): onboarded through the EasyDrama pipeline

German subcorpus (EmDraCor-Ger)

Sources	22 plays from *TextGrid*; 5 plays encoded from scans (*Google Books*); 1 play from GerDraCor (originally from *TextGrid*); 1 play from the *Deutsches Textarchiv*; 1 play from a critical edition (Brauneck, 1975).
Data landscape	Very limited initial availability: only one early modern text (by Gryphius) in GerDraCor at the start of the project. Retrieval further complicated by the wide diffusion of Jesuit Neo-Latin drama in the German-speaking space during the seventeenth century (e.g. Jakob Bidermann's *Cenodoxus*, 1602), which reduced the pool of vernacular candidates.
Issues	/
Gap-filling strategy	Missing titles identified through Alexander (1984) and Meyer (1993); corresponding scanned copies retrieved from Google Books.
Processing	Plays from TextGrid (XML): markup transformation pipeline via Python scripts. Play from the Deutsches Textarchiv (XML): converted to TXT and encoded through the EasyDrama pipeline. Remaining plays (from scans): encoded through the EasyDrama pipeline.

English subcorpus (EmDraCor-Eng)

Sources	28 plays from the newly-created EngDraCor (originally from *EarlyPrint*'s EEBO (*Early English Books Online*) subset; 2 plays directly from *EarlyPrint*'s ECCO (*Eighteenth Century Collections Online*) subset.
Data landscape	Massive text availability through *EarlyPrint*, covering the entire time frame. Stratified sampling was performed on the basis of a master list of all plays in the corpus, which however stopped at 1700 (i.e. included only files from EEBO and not from ECCO). The existing DraCor Shakespeare corpus (ShakeDraCor) was deliberately not included in favour of a broader, non-hypercanonical selection.
Issues	Before being used for sampling, the *EarlyPrint* master list of plays required manual cleaning due to the presence of foreign-language texts and translations.
Gap-filling strategy	Two plays for 1701–1710 randomly sampled from titles listed in the relevant Wikipedia category pages, then manually downloaded from the ECCO collection.
Processing	All *EarlyPrint* texts (XML): markup transformation to DraCor standards, additional curation.

Spanish subcorpus (EmDraCor-Spa)

Sources	17 plays from EMOTHE; 3 plays from CalDraCor; 4 plays from open-access scholarly editions;[1] 6 plays encoded from scans (*Google Books*, *Internet Archive*).
Data landscape	DraCor's Spanish-language collections are limited to nineteenth-century drama (SpanDraCor) and Calderón's full works (CalDraCor). The otherwise useful *Biblioteca Virtual Miguel de Cervantes* offers no direct XML download and its metadata tagging is often imprecise.
Issues	Dating is particularly complex in Spanish early modern drama (e.g. the chronology of Lope de Vega's opus[2]); all dates were cross-referenced against multiple scholarly sources, prioritising the earliest attested year (usually the composition year). Scene segmentation is often unclear in these plays and required particular attention.
Gap-filling strategy	Missing titles identified through the CATCOM (*Catálogo de autores teatrales del siglo XVII*) database (Ferrer Valls, 2013) and other bibliographic resources.
Processing	EMOTHE (XML): markup transformation pipeline via Python scripts. Plays from CalDraCor (XML): additional curation. Scholarly editions (PDF): converted to TXT and onboarded through the EasyDrama pipeline. Remaining plays (TXT, from scans): onboarded through the EasyDrama pipeline.

1 More specifically, from the BIADIG (*Biblioteca Áurea Digital*) collection of the *Grupo de Investigación Siglo de Oro* (GRISO): https://hdl.handle.net/10171/34258.

2 For an attempt to tackle this problem via stylometric methods see Cuéllar (2023).

Italian subcorpus (EmDraCor-Ita)

Sources	12 plays from ItaDraCor; 14 plays encoded from scans (*Google Books*, *Internet Archive*); 3 plays from open-access scholarly editions;[3] 1 play from *Wikisource*.
Data landscape	Many decades were completely unrepresented; in general, the scarce availability of texts in TEI-XML format required substantial encoding efforts from scratch.
Issues	Exclusion criteria: libretti and texts entirely written in local dialects (plays mixing different languages for stylistic reasons, such as Bonicelli's *Pantalone bullo*, were accepted).
Gap-filling strategy	Missing titles identified through historical bibliographies, such as Leone Allacci's *Drammaturgia* (in its revised 1755 edition), and modern ones, such as Herrick (1964).
Processing	Plays from ItaDraCor (XML): additional curation. Remaining plays (TXT, often from scans): onboarded through the EasyDrama pipeline.

3 More specifically, from the *Archivio del Teatro Pregoldoniano* project (https://www.usc.gal/goldoni/biblio/).