

Vocabolario di scrittura umanistica proveniente dalle lettere di Isabella d'Este e Lucrezia Borgia



Lo Scopo

- Rendere fruibile in formato strutturato il lessico della corrispondenza manoscritta di Isabella d'Este e Lucrezia Borgia, figure centrali della cultura rinascimentale italiana.
- Supportare la revisione delle trascrizioni automatiche di manoscritti rinascimentali.
- Destinato alla correzione dei copialettere di Isabella d'Este e delle lettere originali di Lucrezia Borgia

Le Fonti

- Corpus complessivo composto da **297 lettere, 204 immagini e 2.213 righe di testo**.
 - **200 lettere** estratte dai **copialelettere di Isabella d'Este**, provenienti dall'**Archivio di Stato di Mantova** nelle buste 2991, 2993, 2994, 2995, 2996, 2997, 2998 e 3000.
 - **97 lettere originali di Lucrezia Borgia**, conservate nelle buste archivistiche 002 e 003.
- Documentazione tra il **1490 e il 1537**.

L'elaborazione

- A partire dalle trascrizioni prodotte dall'AI delle lettere è stata **generata una lista di vocaboli**.
- **Ogni lemma è stato supervisionato** e sono state **incluse le sue varianti** per aumentare la copertura lessicale.

Trascrizione automatica

- Lo strumento **open** allo stato dell'arte identificato, TrOCR, seppur performante, poco si adatta all'analisi di scritture specifiche.
- Si è sviluppata una **piattaforma** per aiutare gli esperti di dominio nella creazione di un dataset annotato, base che si è utilizzata per il **fine-tuning** dello strumento sulla specifica **scrittura umanistica**.

Entrambi gli strumenti sono a disposizione fra i servizi richiedibili all'interno di DPaaS:

- Piattaforma: **HCR_DataPreparation**
- Modello specializzato: **TrOCR_umanistica**

Il Vocabolario

- Ogni vocabolo è stato sottoposto ad **analisi grammaticale**, con identificazione della **categoria lessicale** e delle **relative caratteristiche** (genere, numero, tempo, modo, ecc.).
- Le **abbreviazioni** sono state **sciolte e normalizzate** nella loro forma estesa, attraverso la consultazione del **Dizionario di Abbreviature Latine ed Italiane di Adriano Cappelli**, nella sua edizione più aggiornata (edizione 7).
- È stata applicata una **normalizzazione minima** relativa ad apostrofazione, accentazione, uso delle maiuscole e minuscole e segmentazione delle parole (unione o separazione).
- Tutte le **altre caratteristiche linguistiche** sono state **conservate nella forma antica attestata nelle fonti**, al fine di **preservare le peculiarità del testo originale**.

Il Dataset in numeri

- 297 lettere
- 204 immagini
- 2.213 righe
- più di 4100 vocaboli annotati grammaticalmente e corredati dai riferimenti ai documenti di origine