Project Proposal: Improving Renaissance Spanish OCR with CNN-RNN Hybrids and Weighted Learning

Kate O'Reilly | katejudithoreilly@gmail.com | https://github.com/kaoreill

Personal Background

Academic Background
My name is Kate O'Reilly, and I'm 21 years old and from Ireland. I study computer science, linguistics, and Spanish at Trinity College Dublin. I'm currently completing an Erasmus year at the Universidad Complutense de Madrid (Spain).

Why this Project?
https://humanai.foundation/gsoc/2025/proposal_OCR1.html

As a student passionate about language, technology, and historical texts, this project feels like a perfect match. My academic focus has been shifting increasingly toward natural language processing, AI, and applied linguistics, and I've always been fascinated by the challenge of bridging past and present through technology.

Transliteration of Renaissance-era Spanish texts desperately needs better tooling. The study is rich in linguistic and historical insight but remains largely inaccessible because modern Optical Character Recognition systems aren't adequately equipped to handle its unique features—like obsolete symbols, archaic letterforms, and irregular layouts.

Working on this project allows me to combine my academic interests and contribute to a real-world problem that has significant value for scholars, linguists, and digital archives. It's exciting to imagine building something that can help unlock centuries-old texts and make them searchable, readable, and usable in modern digital spaces.

Why this Approach?

Tools like Adobe OCR fail on early prints due to:

- Ligatures (e.g. /ſt/, /œ/)
- Obsolete diacritics (e.g. tilde over vowels for nasalization: /ã/)
- Non-standard glyphs (e.g. long /ſ/ vs short /s/)
- Degraded ink/paper

Synopsis

This project describes the development of a free, open-source OCR system for 17th-century Spanish printed texts using weighted CNN-RNN architectures, achieving ≥80% character accuracy (CER ≤20%).

Why this Work Matters

HumanAI and the RenAIssance Project

Most existing OCR models are optimized for modern, standardized fonts in high-quality scans. These models fail dramatically when applied to Renaissance texts, which often include degraded print, ligatures, non-standard orthography, and typographical inconsistencies.

This project contributes directly to the RenAIssance mission: enabling AI systems to understand and transcribe historically rich but digitally underserved materials. Our approach focuses on integrating **historical context** and **linguistic awareness** into the core of the model, something that modern OCR engines don't manage.

Scholars and Digital Archives

Improving OCR for Renaissance Spanish texts means empowering researchers to analyze language change, cultural history, and literary form without needing to transcribe all of the text manually. An improved OCR would mean making rare books more accessible, searchable, and analyzable, helping scholars, librarians, and educators preserve and share cultural heritage.

For digital humanists, archivists, and corpus builders, this work offers tools that go beyond generic OCR, tools that holistically understand the specific challenges of early print.

Technical Approach

Loss Function Design

- **Primary Loss**: **Weighted Cross-Entropy Loss** to penalise misclassifications of rare symbols (e.g., *ç*, *ſ*, *ñ*).
- **Secondary Loss**: **Focal Loss** to focus training on hard-to-classify characters

Handling Extreme Class Imbalance

- **Smoothing**: Add Laplace smoothing (α =1e - 5) to avoid infinite weights for unseen symbols
- **Contextual Rules**: Hardcode positional logic for ambiguous symbols (e.g., replace *ſ* with *s* only in non-final positions)

Data Augmentation for Rare Symbols

- **Synthetic Oversampling**: Use generative scripts to increase rare symbol frequency artificially

Evaluation Protocol

- **Primary Metrics**: CER and WER (computed post-training).
- **Symbol-Specific Accuracy**: Track precision/recall for rare symbols (e.g., *ç*).

Planned Phases

Phase 1: Preparatory Work

- Conduct a deep review of existing OCR architectures (CNN-RNN hybrids)
- Explore and review Renaissance-era Spanish OCR datasets
- Research weighted learning and class imbalance techniques (e.g., weighted cross-entropy, and class-aware sampling)
- Document common rare symbols, ligatures, and diacritics from 17th-century texts

Phase 2: Model Implementation

- Build baseline OCR model
- Apply weighted learning to underrepresented character classes
- Train and evaluate on a historical Spanish dataset
- Metrics: Character Error Rate (CER), Word Error Rate (WER)

Phase 3: Enhancement and Fine-Tuning

- Augment data to simulate printing degradation and rare symbol appearances
- Tune weights and loss functions (e.g. CER and WER)
- Compare against baseline and evaluate improvements
- Test on unseen texts and document generalisation performance

Final Deliverables

- Trained model and inference scripts
- Evaluation report and graphs
- Codebase with documentation
- Sample transcriptions on historical Spanish texts
- README and contribution notes

Timeline (175-hour Project)

Hours	Phase	Tasks
25hrs	Data Prep	- Clean/organise provided images/transcripts (**10hrs**) - Split into train/val/test (**5hrs**) - Analyze symbol frequencies for class weights (**10hrs**)
25hrs	Baseline Model	- Implement CNN + BiLSTM + CTC loss in PyTorch (**15hrs**) - Train initial model (**8hrs**) - Compute CER/WER baseline vs Tesseract (**2hrs**)
30hrs	Weighted Training	- Add focal loss + symbol weights (**10hrs**) - Train with class balanced sampling (**15hrs**) - Validate rare symbol recall (**5hrs**)
15hrs	Basic Augmentation	- Add synthetic ink bleed/paper texture (**10hrs**) - Train with augmented data (**5hrs**)
30hrs	Fine-Tuning	- Layer pruning (reduce CNN filters 20%) (**10hrs**) - Regional rules (**5hrs**) - Final training cycle (**15hrs**)
20hrs	Evaluation	- Test CER/WER on unseen texts (**10hrs**) - Regional bias analysis (**5hrs**) - Error analysis and visualizations (**5hrs**)
15hrs	Documentation	- Code comments/README (**5hrs**) - User guide: installation and inference (**5hrs**) - Contribution guidelines (**5hrs**)
15hrs	Deployment Prep	- Export trained weights (**2hrs**) - Create sample transcriptions (**8hrs**) - Finalise GitHub repo (**5hrs**)
Total: 175h		

Ethical Considerations

Regional/Dialectal Bias

Renaissance-era Spanish texts vary significantly by region (e.g., Castilian vs. Andalusian prints) due to historical linguistic diversity. An OCR model trained on Castilian-dominated corpora may systematically fail on texts from other regions, erasing dialectal nuances and perpetuating cultural marginalization.

Poor performance on non-Castilian texts could skew linguistic studies, reinforcing the false narrative of Castilian as the "standard" Renaissance Spanish.

Mitigation

- Curate regionally balanced training data
- Add region-specific mappings (e.g., *ss* → *s* in Andalusian, *x* → *j* in older Castilian)
- Disclaimers such as "This model performs best on Castilian prints; Andalusian texts may require manual verification."

About Me

I am a third-year computer science student at Trinity College Dublin, specializing in **AI, NLP, and Spanish linguistics**, with hands-on experience building OCR tools during my Erasmus year at Universidad Complutense de Madrid (Spain). My unique qualifications include:

- **AI/ML Development:** Proficient in Python and PyTorch, with coursework in neural machine translation and transformer architectures.
- **Software Engineering:** Led Agile development of a Java-based Blackjack game at UCM using Scrum and UML modeling, skills directly applicable to structuring this OCR pipeline.
- **Applied Linguistics:** B2 Spanish certification (and living in Madrid, immersed in the language for the last 8 months) + philological training in historical orthography through UCM's literature courses

I have received high grades in these courses, such as a 9.0/10 in Historia de la Lengua Inglesa (History of the English Language, a historical linguistics course in UCM), an 8.6/10.0 in Artificial Intelligence (a course specifically on Machine Learning in Python), and a 7.4/10.0 in Software Engineering (with emphasis on code maintainability).

My dual training in computational methods (Trinity College Dublin) and Spanish paleography (UCM) allows me to tackle both technical and linguistic challenges in this project.

Glossary

Term	Definition
OCR	Optical Character Recognition: Extracting text from images
CNN	Convolutional Neural Network: Detects spatial patterns

RNN	Recurrent Neural Network: Models sequential data (e.g., text)
LSTM	Long Short-Term Memory: RNN variant for long sequences
CER	Character Error Rate: (Insertions + Deletions + Substitutions) / Total Characters
WER	Word Error Rate: Same as CER but at word level
Ligature	Joined characters e.g., /æ/
Diacritic	Accent marks, e.g., /ċ/