

# An Unsupervised Model of Orthographic Variation for Historical Document Transcription

Dan Garrette

Computer Science & Engineering  
University of Washington

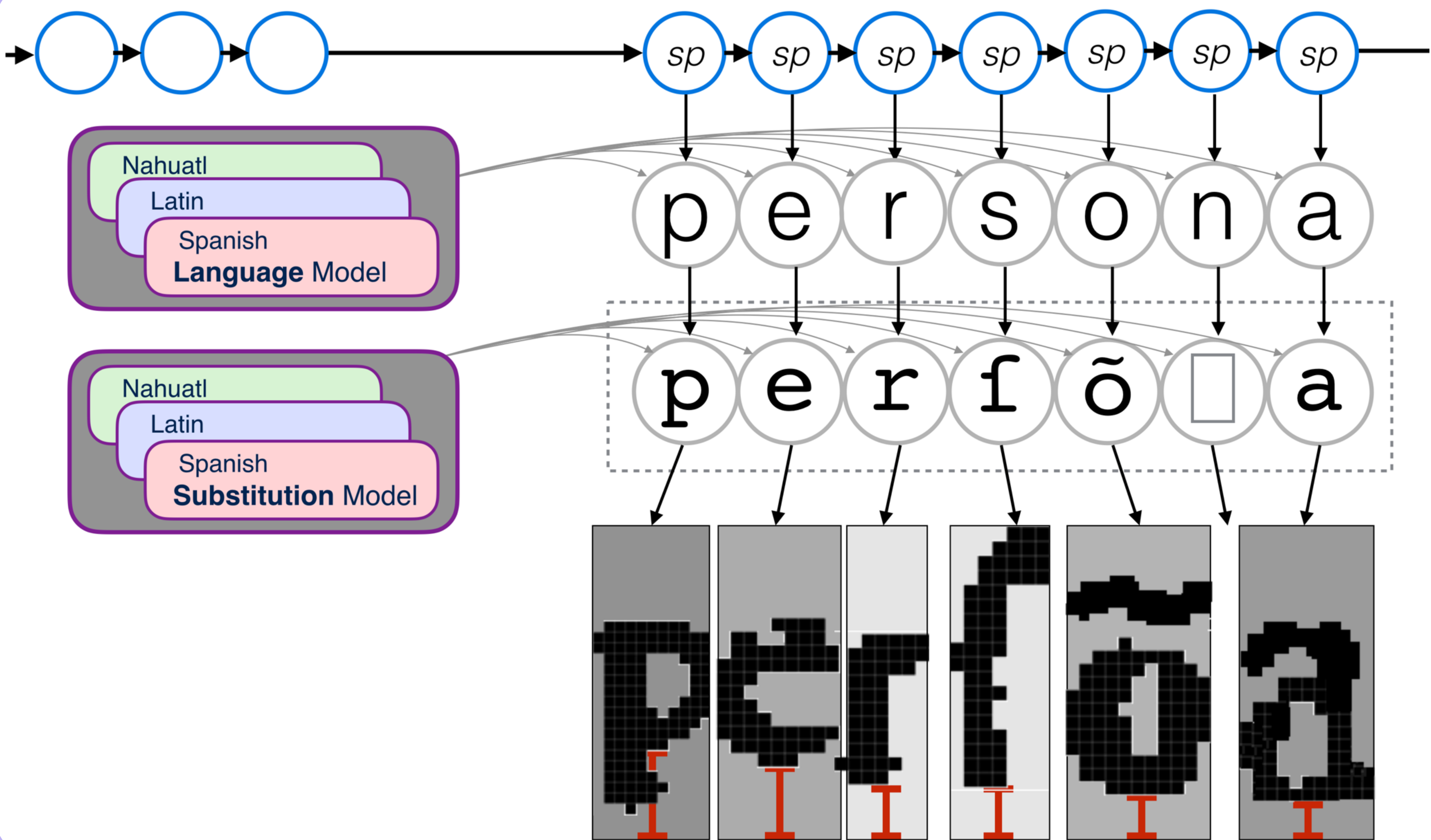
Hannah Alpert-Abrams

Comparative Literature  
University of Texas at Austin

## Generative Model

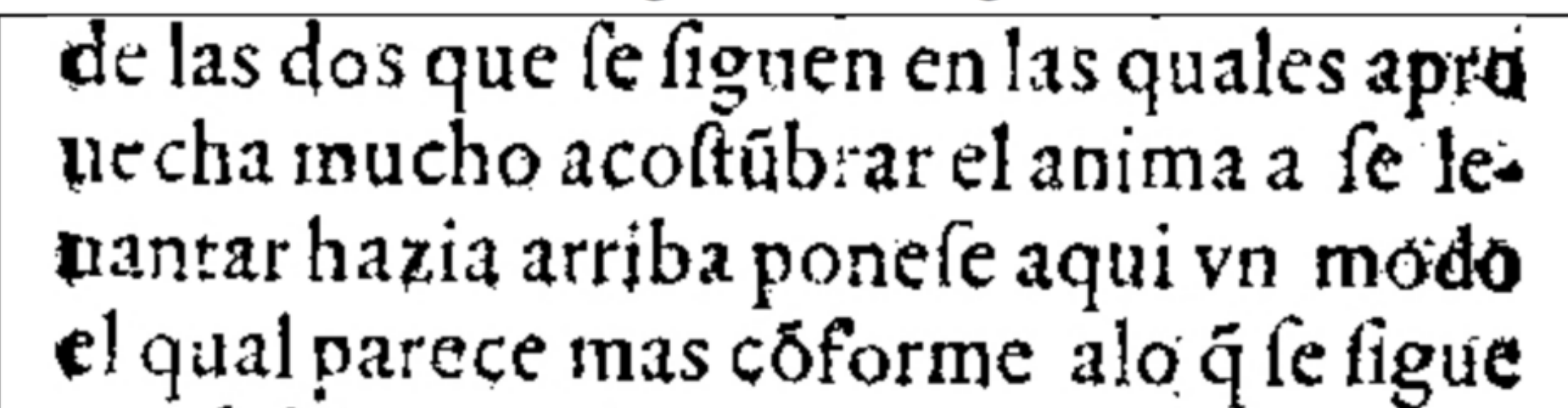
For each character state, we generate: 1) **language**, 2) **modern-orthography character**, 3) **printed character**, and then 4) **typesetting**. Our system jointly outputs two transcriptions: one *literal* (standard OCR output), and one *automatically normalized* (for clean results).

Our model is embedded within the Ocular OCR system of Berg-Kirkpatrick, Durrett, and Klein (2013)



## System Output

### Original Image



### Diplomatic

de las dos que se figuen en las quales apro  
uecha mucho acostũbrar el anima á se le-  
pantar hazia arriba ponefe aqui vn modo  
el qual parece más cõforme á lo q̃ se figue

### Normalized

de las dos que se siguen en las quales  
aprovecha mucho acostumbrar el ánima á se  
levantar hacia arriba pónese aquí un modo  
el cual parece más conforme á lo que se sigue

## Results

Orthographic variation strategy	Diplomatic		Normalized	
	CER	WER	CER	WER
No handling	13.2	45.7	17.4	47.6
Hand-written rules	<b>8.5</b>	<b>30.8</b>	13.1	37.9
Unsupv. joint model	8.6	32.7	<b>9.5</b>	<b>27.6</b>

## Learned Substitution Probs

<i>c</i>	<i>g</i>	$freq(sp., c, g)$	$P_{spanish}^{GLYPH}(g   c)$
-	ELIDED	52	0.0881
ó	o	31	0.0526
s	f (long s)	325	0.0352
q	q̃	9	0.0222
n	ELIDED	57	0.0136
v	u	55	0.0129
o	õ	20	0.0091
c	cc	23	0.0028

	char sub. (c → q)	char sub. (s → long s)	elision (que → q̃)	accent drop (ó → o)	doubled (c → cc)	typo (e → r)
Original image	qual	<b>esta</b>	<b>aql</b>	confideracion	peccados	<b>Primeramrnte</b>
Baseline trans.	qual	eña	á ol	confideracion	peccados	Primeraminte
Our diplomatic trans.	qual	esta	<b>aql</b>	confideracion	peccados	Primeramrnte
Our normalized trans.	cual	esta	<b>aquel</b>	consideración	pecados	Primeramente