An Unsupervised Model of Orthographic Variation for Historical Document Transcription

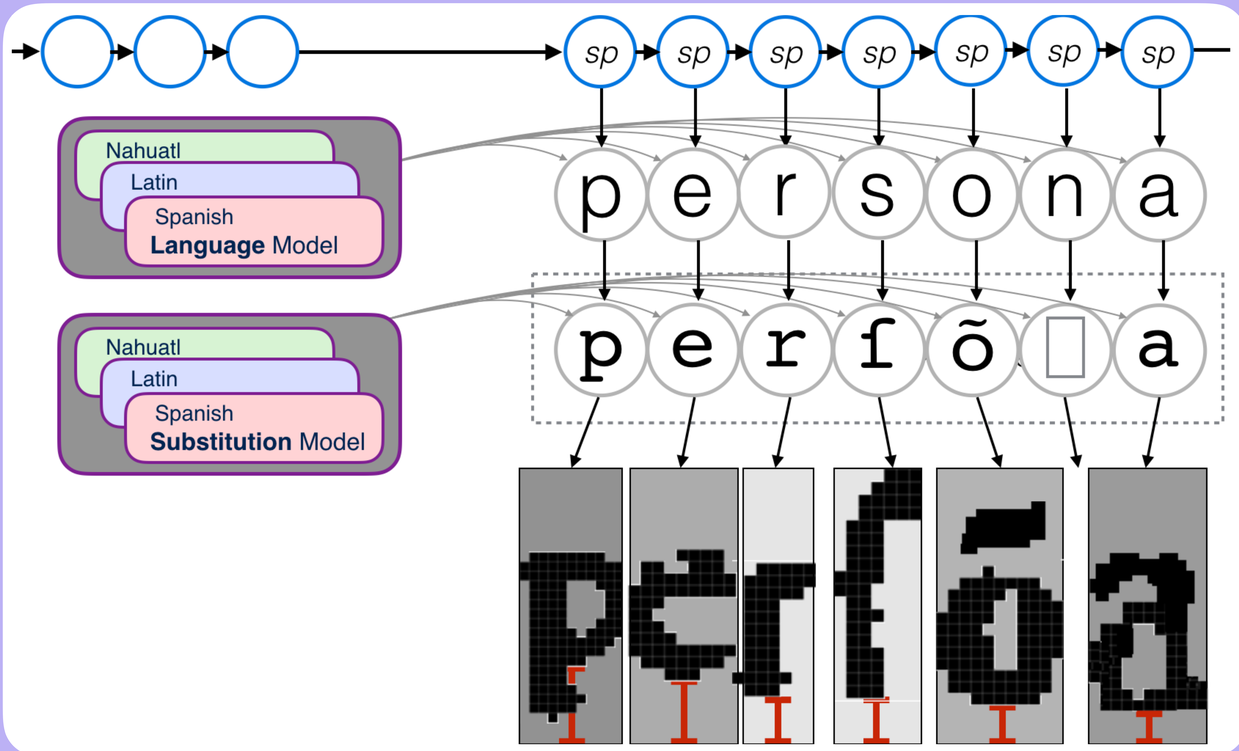
Dan Garrette
Computer Science & Engineering
University of Washington

Hannah Alpert-Abrams
Comparative Literature
University of Texas at Austin

Generative Model

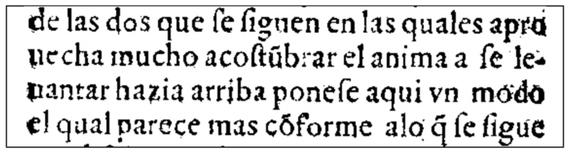
Our system jointly models two transcriptions: one *literal* and one *automatically normalized*. For each character state, we generate: 1) **language**, 2) **modern-orthography character**, 3) **printed character**, and then 4) **typesetting**.

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



System Output

Original Image



Output - Literal

de las dos que **ſe** **ſ**iguen en las quales apro
uecha mucho aco**ſtũ**brar el **a**nima á **ſ**e le-
pantar ha**z**ia arriba **p**one**ſe** aqui **v**n modo
el **q**ual parece más **cõ**forme á lo **q̃** **ſ**e **ſ**igue

Output - Normalized

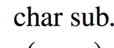
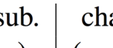
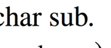
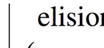
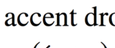
de las dos que **se** **s**iguen en las quales
aprovecha mucho aco**stum**brar el **á**nima á **se**
levantar ha**c**ia arriba **p**ónese aquí **u**n modo
el **c**ual parece más **con**forme á lo **que** **s**e **s**igue

Results

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

Learned Substitution Probs

c	g	$freq(sp., c, g)$	$P^{\text{GLYPH}}_{spanish}(g \mid c)$
-	ELIDED	52	0.0881
ó	o	31	0.0526
s	ſ (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						
Baseline trans.	qual	eña	á ól	confideracion	peccados	Primeramınte
Our diplomatic trans.	qual	eſta	aq̃l	confideracion	peccados	Primeramrnte
Our normalized trans.	cual	esta	aquel	consideración	pecados	Primeramente