

Optical Character Recognition

Hannah Alpert-Abrams
University of Texas at Austin

Slides:
halperta.com/mla17.pdf

chriſtiana. Fo 10.
 ¶ Nic. ca ypampa inic yehuantin quicui
 lozque ininemilztzin ynictlalticpacmo
 nemitico in totecuyo Jeſu xp̃o: intleyn
 quimochihuilico in quimotemachtilico:
 inicttehuantin tictotepotztoquilizque y
 nitemachtiltzin. ¶ No. Catlehuatl ino
 quicuiloque in nahuintin in Euangelis-
 tame.
 ¶ Nic. Ca yehuatl ynauhtlamantli
 yn Euangelio ynipan mopia ynixquich
 totech monequi in ticneltocazque yn tic
 chihuazq̃. Yhuā intleyn tictlalcahuizq̃
 intictelchihuazque. Yhuan intley tichi
 huazq̃ yye yxq̃ch totech monequi. Tluh
 yntlacamo ticneltocazque in tlein qui-
 cuilotiaque çan nimā ahueltitomaquix-
 tizqui.
 ¶ No. nicmatiznequi ynictiazque yn il-
 huicac cuix ça yeyyo ticneltocazque yn
 oquicuilotiaque.
 ¶ Nic. ca itlacamo tictopielicā tictone
 miliztican initeotenahuatiltzin, yyehuā
 tin inquicauhtiaque caniman ahuel tiaz
 que yn ilhuicac.

b ij

chriſtiana. Fo 10.
 ¶ Nic. ca ypampa inic yehuantin quicui
 lozque ininemilztzin yn ictlalticpacmo
 nemitico in totecuyo Jeſu x̃~po: intleyn
 quimochihuilico in quimotemachtilico:
 inicttehuantin tictotepotztoquilizque y
 nitemachtiltzin. ¶ No. Catlehuatl ino
 quicuiloque in nahuintin in Euangelij-
 tame.
 ¶ Nic. Ca yehuatl yn auhtlamantli
 yn Euangelio yn ipan mopia yn ixquich
 totech monequi in ticneltocazque yn tic
 chihuaz~q. Yhu~a intleyn tictlalcahuiz~q
 intictelchihuazque. Yhuan intley tichi
 huaz~q yye yx~'qch totech monequi. Tluh
 yntlacamo ticneltocazque in tlein qui-
 cuilotiaque çan nim~a ahueltitomaquix-
 tizqui.
 ¶ No. nicmatiznequi yn ic tiazque yn il-
 huicac cuix ça yeyyo ticneltocazque yn
 oquicuilotiaque.
 ¶ Nic. ca ~itlacamo tictopielic~a tictone
 miliztican initeotenahuatiltzin, yyehu~a
 tin inquicauhtiaque caniman ahuel tiaz
 que yn ilhuicac.

b ij

christiana. Fo 10.

¶ Nic. ca xpampa inic yehuantin quicui
 lozque ininemiztzin ynictlalticpac mo
 nemitico in totecuvo Jesu xpo: intle yn
 quimochihuilico in quimotemachtilico:
 inictehuantin tictotepotztoquilizque y
 nitemachtiltzin. **¶** Mo. Catlehuatl ino
 quicui loque in nahuintin in Euangelis-
 tame.

¶ Nic. La yehuatl ynauhtlamantli
 yn Euangelio ynipan mopia ynirquich
 totech monequi in ticneltocazque yn tic
 chihuaazq. Yhua intle yn tictlalcabhuizq
 in tic telchihuaazque. Yhuan intle y tichi
 huaazq yye yrqch totech monequi. Ahu
 yntlacamo ticneltocazque in tlein qui-
 cuilotiaque can nimā ahueltitomaquix-
 tizqui.

¶ Mo. nicmatiznequi ynictiazque yn il-
 huicac cuix ca yeyvo ticneltocazque yn
 oquicui lotiaque.

¶ Nic. ca itlacamo tictopielicā tictone
 miliztican initeotenahuatiltzin, y yehua
 tin inquicauhtiaque caniman ahuel tiaz
 que yn ilhuicac.

b ij

o — M r^ -^ v^SO |>»00 Cv O — «
 r^ -«t- «^lo r-^oo Oso — r^
 vo SO vO \0 vO *0 vO vO nO so \0
 sOvO vososOnOvOvOvosOvOvO

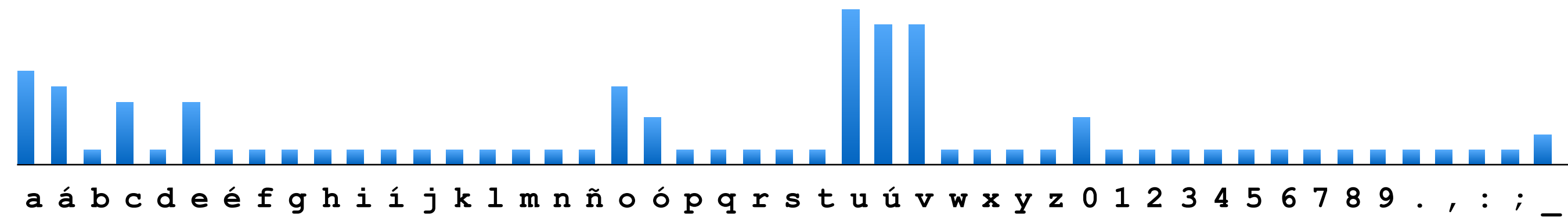
00 (7s O — «*< '◆N ^ w^sO r-»00
 (7s o «- n r^ 't- «^VO f>.00 Cv o
 w^ w^vO sovoovosovovovo>ovo r**
 r^r^i^i^c^r>.c^r>. r**oo

VO r**00 Cv o — *^ «^ -^ «^SO
 t>i30 J\ o — r«

o — M r^ -^ v^SO |>»00 Cv O — «
 r^ -«t- «^lo r-^oo Oso — r^
 vo SO vO \0 vO *0 vO vO nO so \0
 sOvO vososOnOvOvOvosOvOvO

00 (7s O — «*< '◆N ^ w^sO r-»00
 (7s o «- n r^ 't-

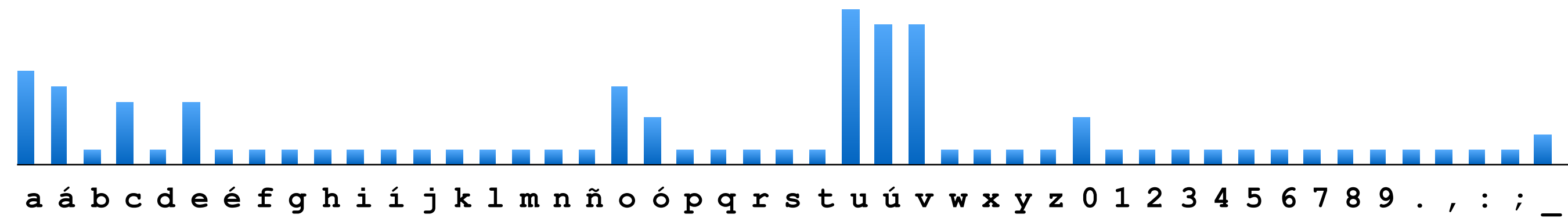
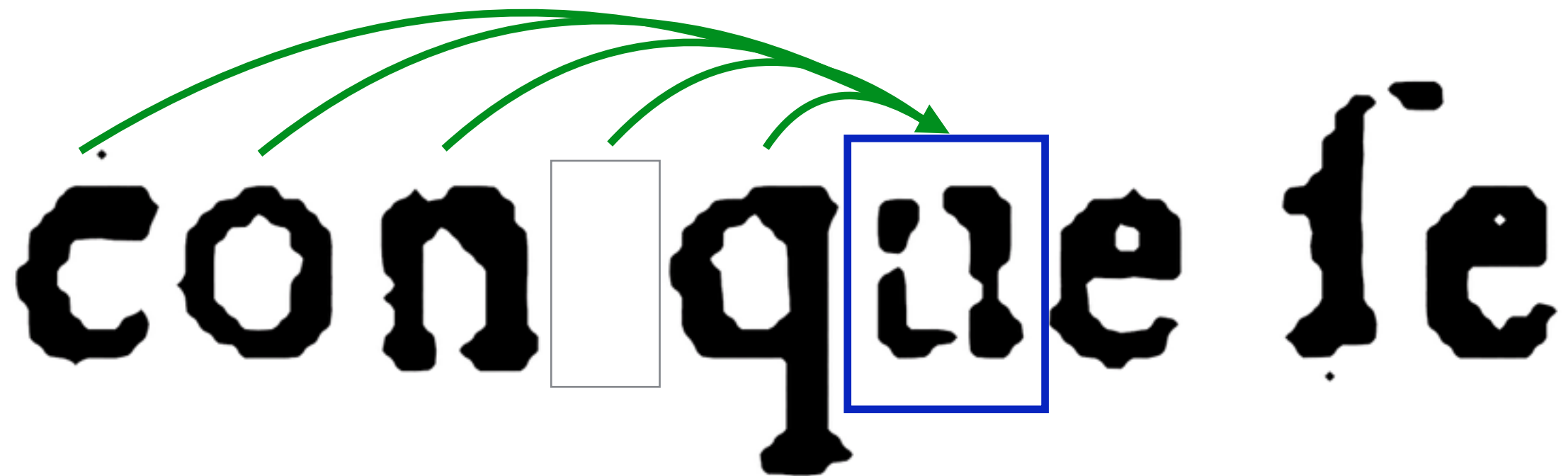
Font Model



OCR works by analyzing the visual characteristics of an image

Language Model

conque le



OCR uses statistical models of what language looks like to recognize difficult letters.

OCR transcribes texts using statistical models of their **visual** and **linguistic** characteristics

Linguistic errors

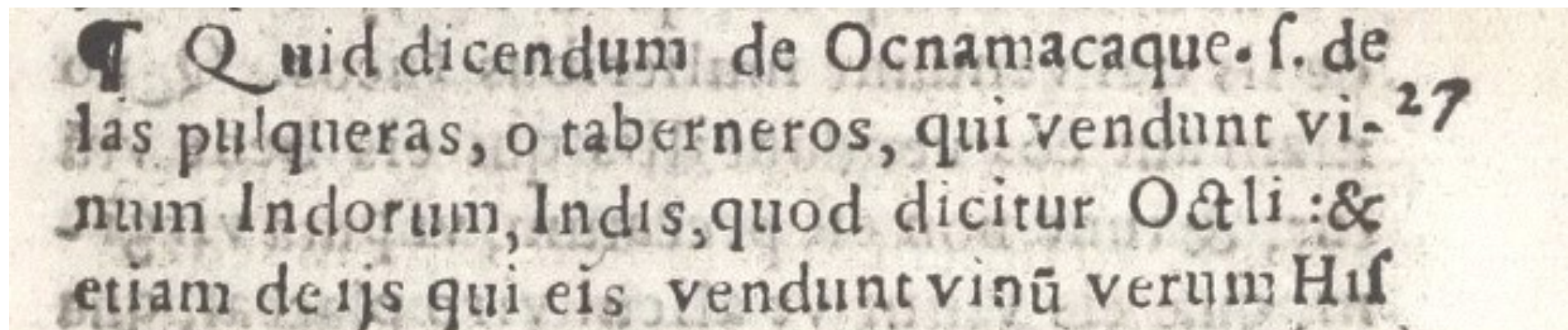
Dios?y haviẽdo de resp. Ca huel imeixtintzi

1. tentations made defensive Carlucci International
2. Uto wry haufe do do to fp: Ca hunt fnicket into a
3. Dios, y haviẽdo derefp. Ca huel imeixtintzi

OCR fails when the language doesn't match its expectations of what language should be.

Linguistic opportunities

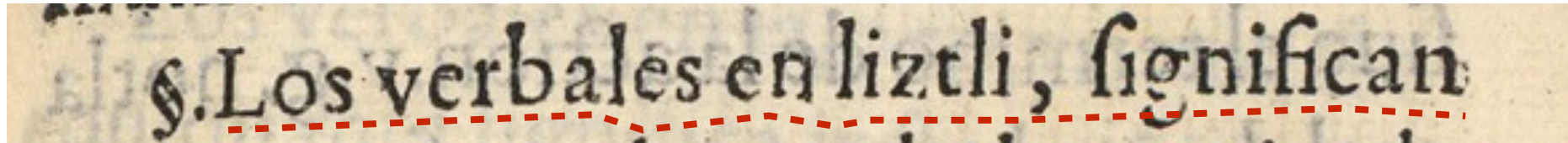
"¶ Quid dicendum de Ocnamacaque. f. de las pulqueras, o taberneros, qui vendunt vinum indorum, indis, quod dicitur Octli .&. etiam de iis qui eis venduntur nō verum Hi



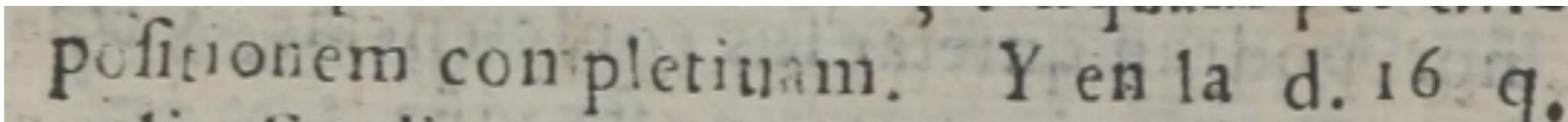
OCR can introduce new kinds of metadata.

Visual errors

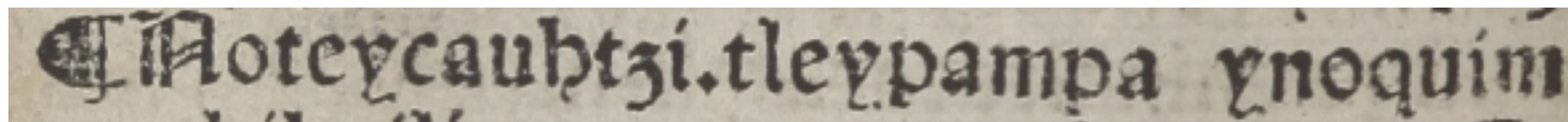
Wandering baseline



Uneven inking

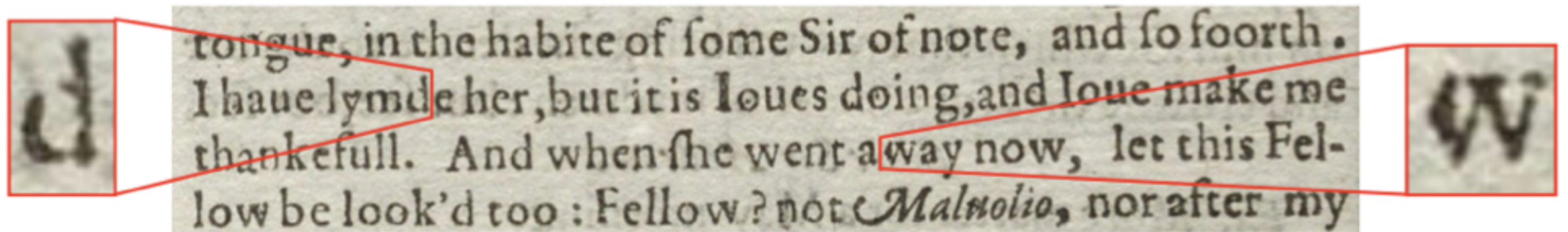


Unfamiliar Typefaces



OCR fails when the images don't match its expectations of what letters should be.

Visual opportunities



OCR can be used as a research tool.

OCR analyzes the visual and linguistic characteristics of a document.

It can produce new kinds of metadata.

It can create new opportunities for research.

Hannah Alpert-Abrams
sites.utexas.edu/firstbooks.