Lyric Alignment on Plainchant Manuscripts

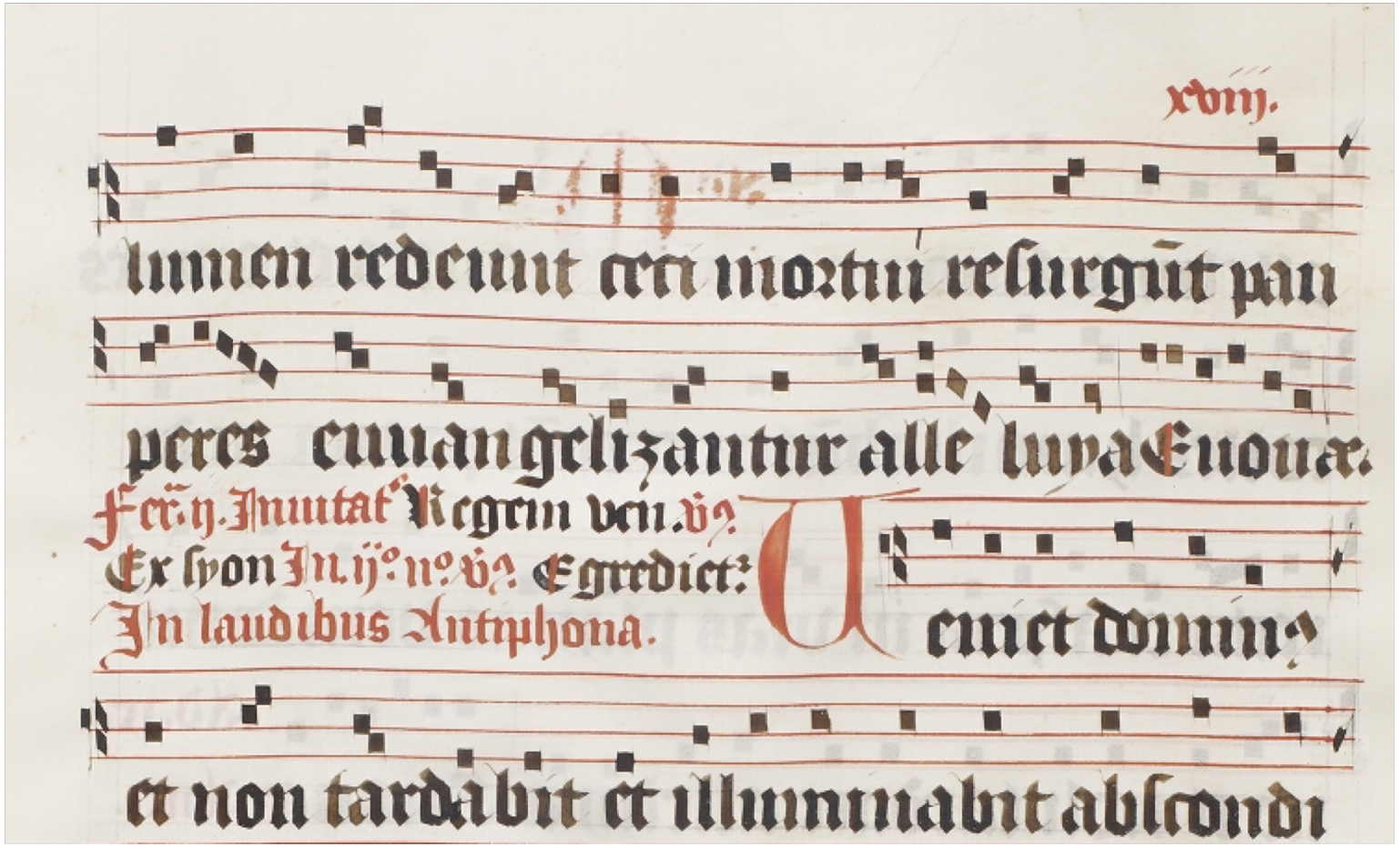
Tim de Reuse

SIMSSA Workshop XVII

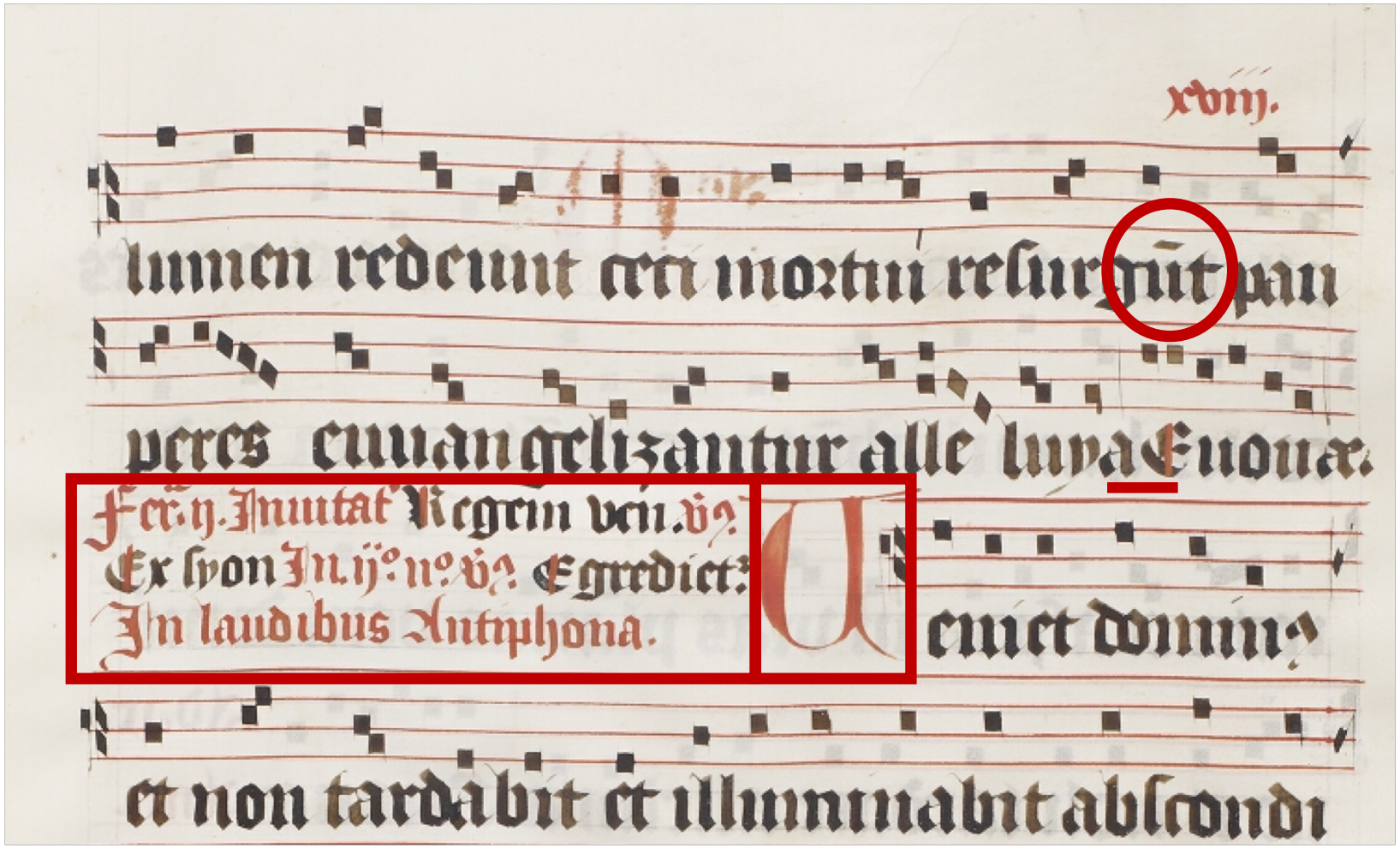
1/12/2018

Associating Neumes to Syllables of Text

- Must locate text first
- Optical Character Recognition (OCR) on handwritten sources is difficult!
- Existing transcripts do not indicate *where* text is on page



Non-musical Text



Abbreviations

Inconsistent Spacing

Ornamental Letters

TRANSCRIPT: lumen redeunt ceci mortui resurgunt pauperes euvangelizantur alleluya Euouae Veniet dominus et non tardabit et illuminabit abscondi (…)

Aligning to Imperfect OCR

- Run OCR on handwritten source
  - Find out where *most* of the characters are
  - Expect lots of errors
- Compare OCR result with correct transcript
  - Correct errors in the OCR and find all characters
  - Identify non-musical text, abbreviations, etc.

Training the OCR

- ~650 lines of text transcribed for training a model
  - 42 Pages of the Salzinnes Antiphonal
  - That's not a lot for handwritten OCR!
  - Using OCRopus open-source OCR system
- ~80% per-character error rate after 24 hours training
  - Good enough!

TRANSCRIPT:
`lumen redeunt ceci mortui resurgunt pauperes euvangelizantur alleluya Euouae`

OCR:
`lmmen redeunt ceci mortiiui resurgūt pan yeres eyugelisantur alle luya Euonae`

Global Sequence Alignment

- Given two sequences, edit them to make them “match up”
  - Cost minimization: as few gaps as possible for as many matches as possible

```
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor
incididunt ut labore et dolore magna aliqua.
```

```
Lorem ipsum dollllllor acsit amet, consectur di.s elit,, eiusmmd tempodsr
incididunt ut lb ore etmagna aliqua.
```

```
Lorem ipsum do_____lor __sit amet, consectetur adipiscing elit, sed do
Lorem ipsum dollllllor acsit amet, consec__tur _di_.s____ elit,______,

eiusmod tempo___r incididunt ut labore et dolore magna aliqua
eiusmmd tempodsr  incididunt ut lb ore et________magna aliqua
```

Aligning the Transcript to OCR Output

TRANSCRIPT:
lumen redeunt ceci mortui resurgunt pauperes euvangelizantur alleluya Euouae

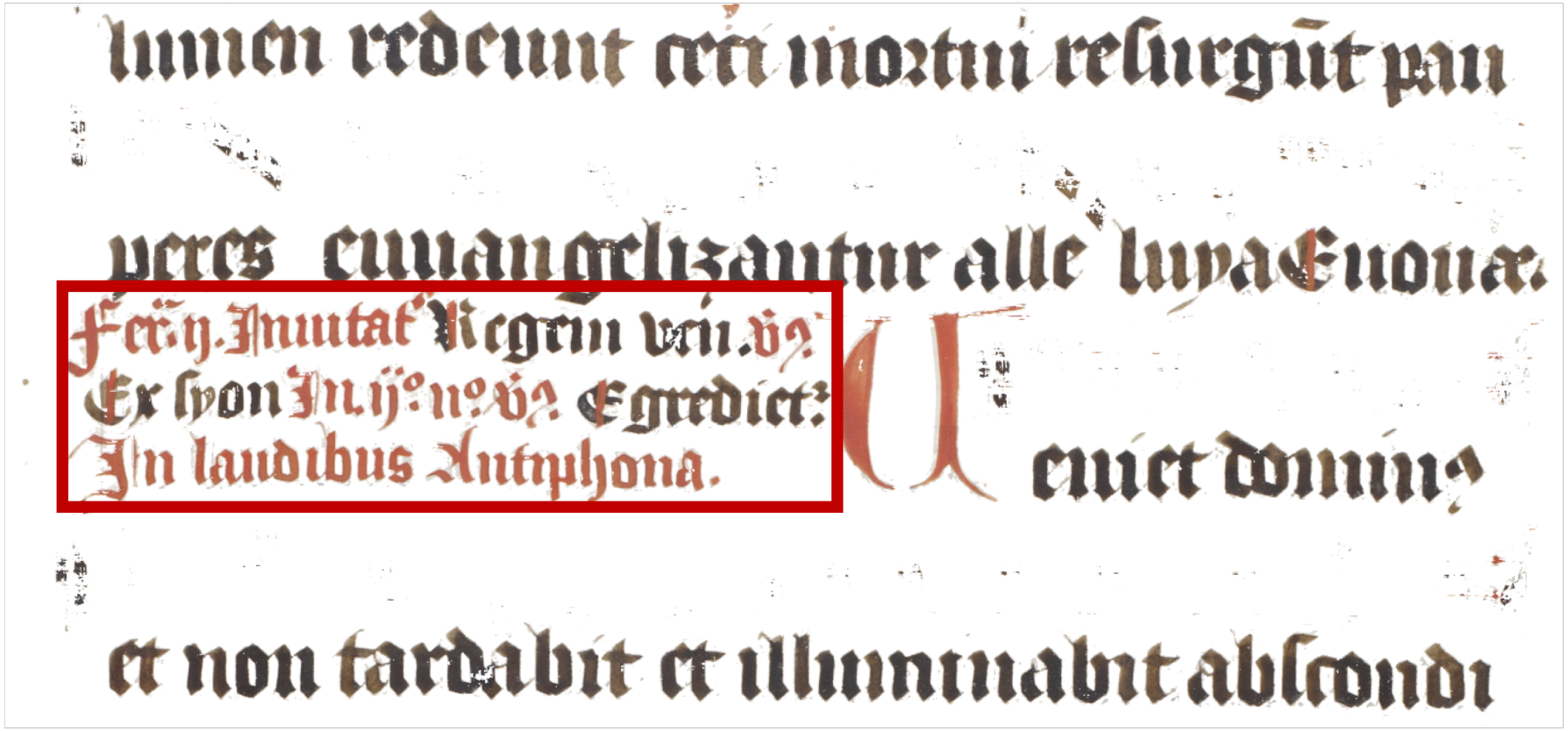
OCR:
lmmen redeunt ceci mortiiui resurgūt pan yeres eyugelisantur alle luya Euonae

ALIGNMENT:

```
lumen redeunt ceci mort__ui resurgunt pauperes
lmmen redeunt ceci mortiiui resurgu_t panyeres

euvangelizantur alle_luya Euouae
e__yugelisantur alle luya Euonae.
```

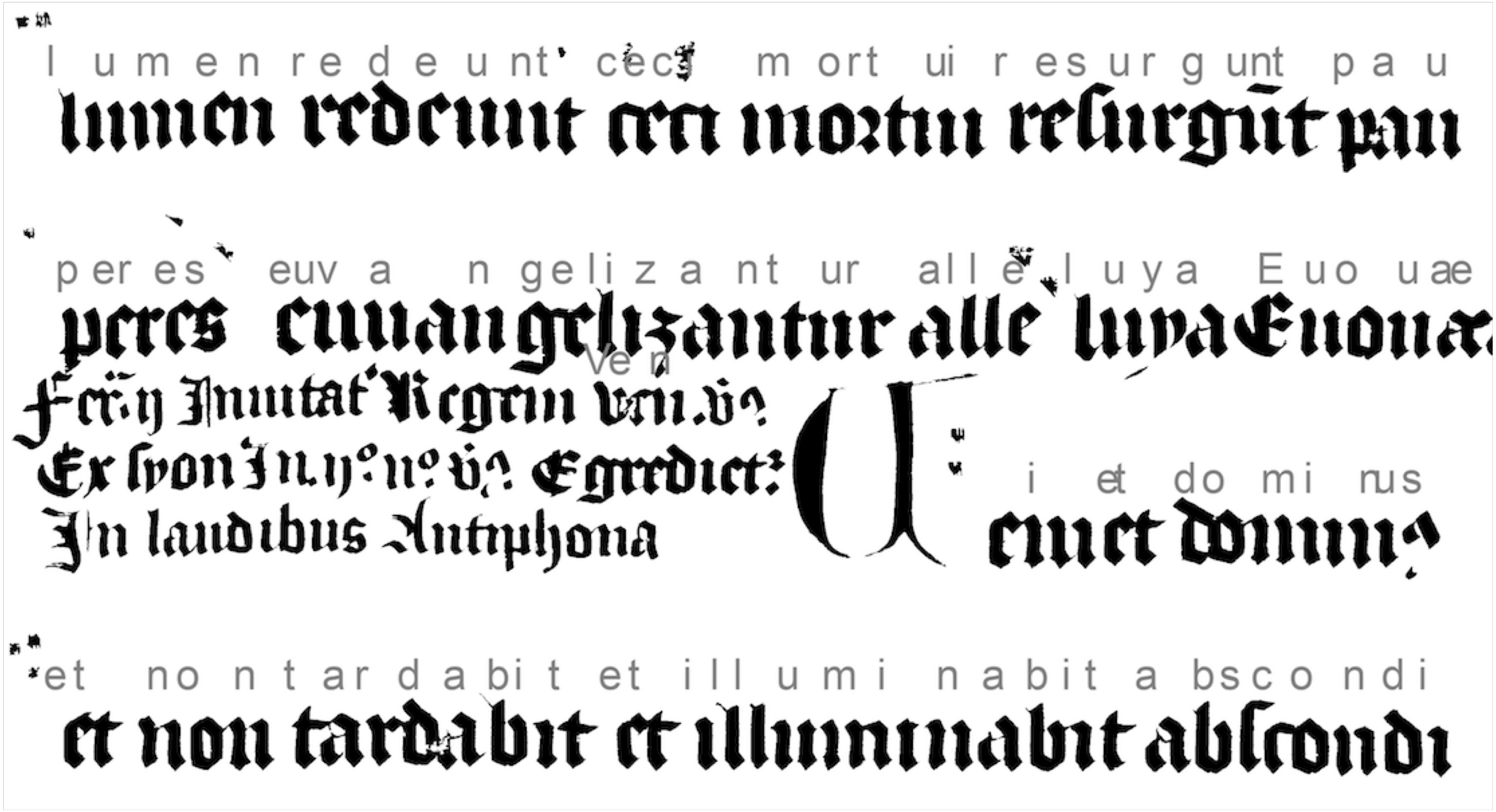
The OCR gets characters wrong, but the incorrect character is often aligned to the correct one in the transcript

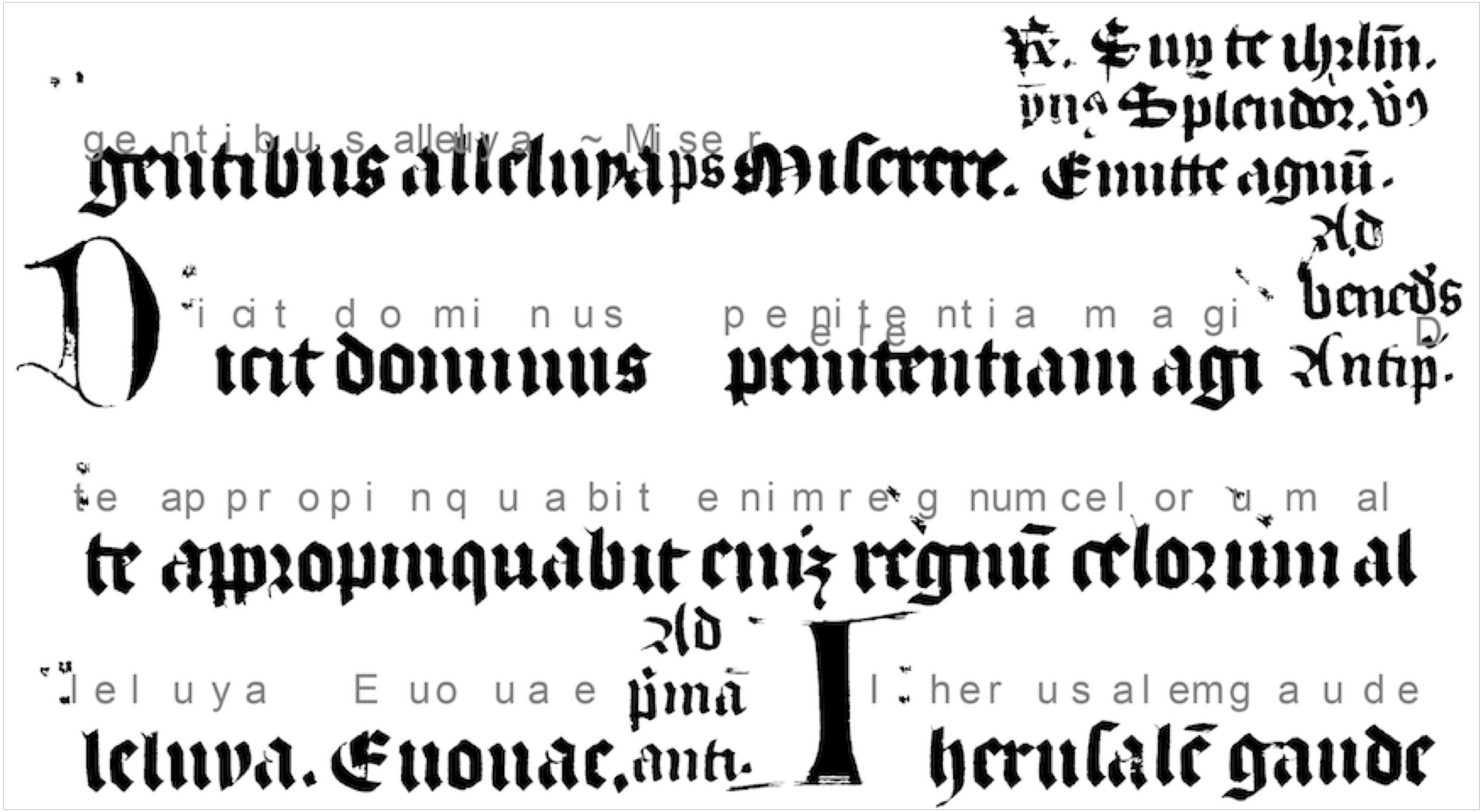


lumen redeunt ceci mort__ui resurgunt pauperes euvangelizantur alle_luya Euouae_____
lmmen redeunt ceci mortiiui resurgu_t panyeres e__yugelisantur alle luya Euonae. cx.n nni

_____ Ven_____iet dominus et non tar
tat segeim ven.vus. In landibus Antiplhona In laud ibus Antiphona emet dominus_et non tar

A long gap: The OCR found non-musical text not present in the transcript, and the sequence alignment identified it as such


l u m e n r e d e u nt cec m ort ui r es ur g unt p a u
per es euv a n geli z a nt ur all e l u y a E uo uæ
Ve n
i et do mi nus
et no n t ar d a bi t et i l l u m i n a bit a bsc o n di


gentibus alleluya ~ Miser
gentibus alleluia Miserere. Emitte agnū.
Dicit dominus penitentiam agi
dicit dominus penitentiam agi
te appropinquabit enim regnū celorum al
te appropinquabit enim regnū celorum al
leluya Euo uae
leluya. Euouae.
her usalemg aude
herusalē gaude

Thank you!