



STATE

Assisted Text Transcription System

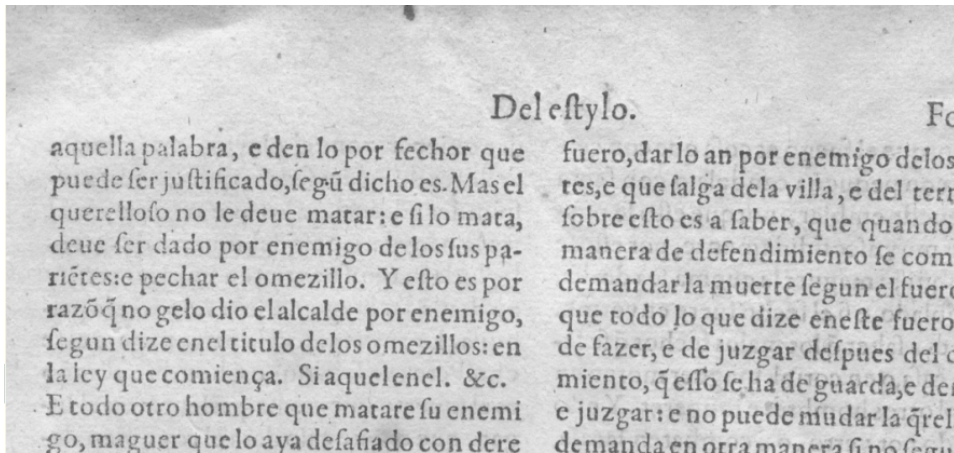
David Llorens, Andrés Marzal, Vicente Palazón, Federico Prat, Juan Miguel Vilar (UJI, Spain)
María José Castro, Salvador España, Joan Pastor, Francisco Zamora (UPV, Spain)

The people



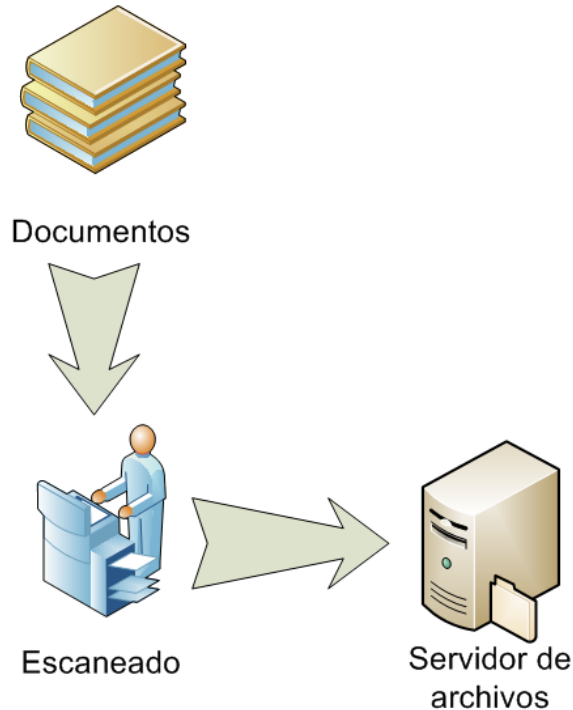
Automatic? Transcription

- ▶ **Optical Character Recognition (OCR)** systems perform **poorly**:
 - ▶ On damaged documents.
 - ▶ On ancient documents.
 - ▶ On handwritten documents.
 - ▶ On text with non-estandar characters or fonts.
- ▶ It is always needed to **supervise and correct** the automatic transcription.



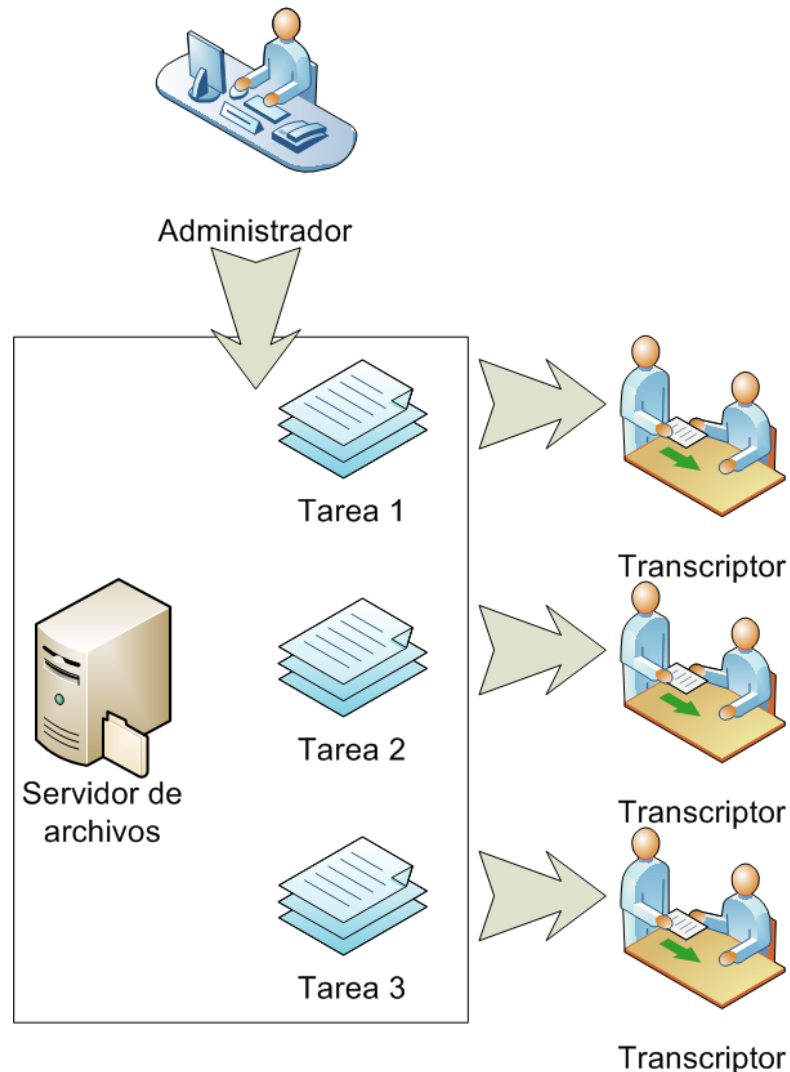
Mr. Dowell finds it easier to take it out of mothers, childrens and sick people than to take on this vast industry," Mr. Bonn commented icily. "let us have a full inquiry into the cost of drugs and the pharmaceutical industry." The health of children today owed much to the welfare food scheme. It was maintained during the war. Now in conditions of Tory affluence it seemed it could not be carried on.

Documentary resources digitization



- ▶ Documents are **scanned...**
- ▶ ...and the obtained images are uploaded to a **file server.**

Documentary resources digitization



- ▶ The **administrator** divides the documents in groups of pages (**tasks**).

- ▶ Each task is assigned to a **human transcriber**.

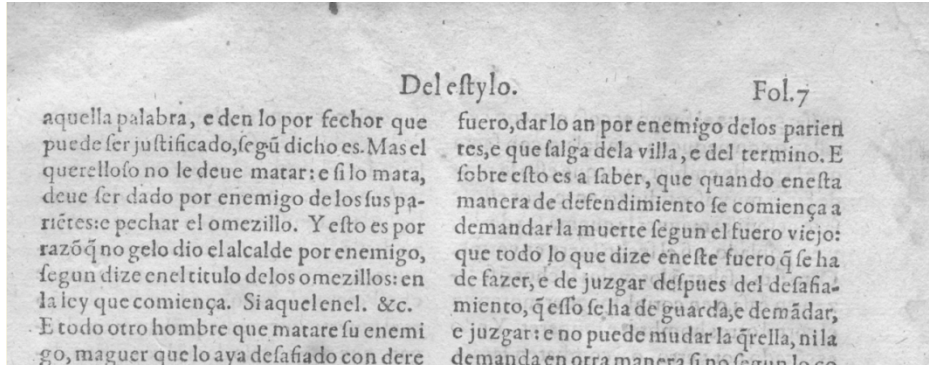
His dutie is to obtain a text file from the original document.

Documentary resources digitization



- ▶ Each document is **transcribed with an OCR system**.
- ▶ The resulting text contains **many errors**.
- ▶ Text editing: the transcriber must **manually correct** the text.
- ▶ It is the most **time consuming stage**, where it is easy to introduce **errors**.
- ▶ Finally, the correct transcription is uploaded to the server.
- ▶ Now, the digital information can be exploited in many ways.

How is text editing performed?



aquella palabra , e den lo por fechor que
.puede serj u [st]i[si]cado,seg[un] dicho es.Mas el
quer.eloso no le deue matar: e si lo mata,
deue ser dado por enemigo de los sus pa-
ri[en]tes:e pechar el omezillo. [ss] e[st]o es por
raz[on] [que] no gelo dio el alcalde por enemigo,
segun dize en el titulo de los omezillos: en
.la ley que comiença. bi aquel enel. ac.
.E todo otro hombre que matare su enemi
go, m.uguer que lo aya desa[si]ado con dere

- ▶ Comparing the page with its transcription is **very uncomfortable and difficult**.
- ▶ The human expert must
 - ▶ **jump** continuously from the image to the text,
 - ▶ **find** each error,
 - ▶ **move** with the mouse or keyboard,
 - ▶ and **type** the correction.

Ancient documents

Problems get worse with ancient documents

- ▶ Stains, handwritten annotations on the margins, bad conditions, cracks, patches, disappearing ink in some regions, and so on:

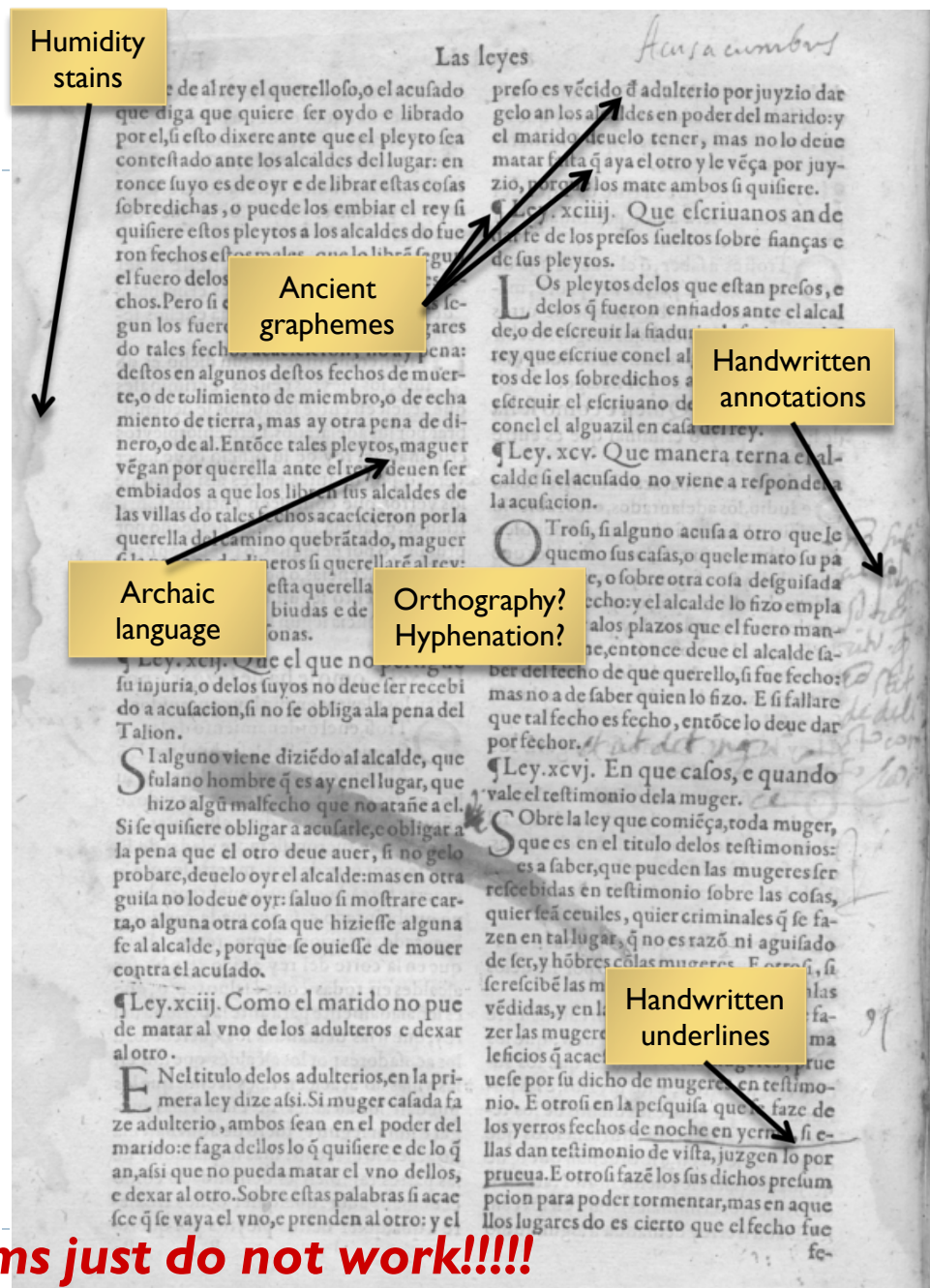
Text extraction is difficult for OCR systems.

- ▶ Special fonts or ancient graphemes:

OCR systems do not recognize them.

- ▶ Non-modern orthography or syntax and archaic language:

Language models of OCR systems are not suitable.

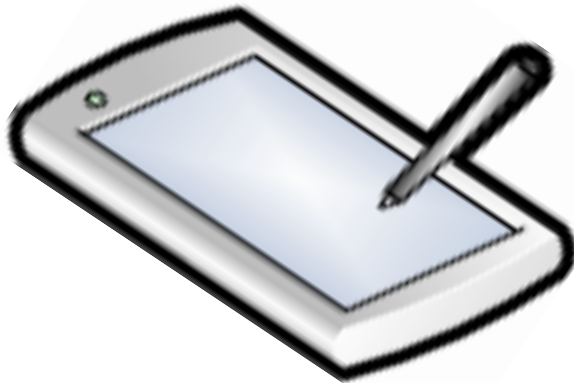


Our aims

- ▶ Our transcription system **integrates advanced and accurate tools**
 - ▶ for *image processing*,
 - ▶ to *detect page layout*,
 - ▶ for *text recognition*.
- ▶ **But also... it assists the human expert to**
 - ▶ *partially automatize* the text editing process, and
 - ▶ attend to usability in order to ease the *interactive text editing process*.
- ▶ And the recognition system includes **adaptive learning**: it learns from samples of each new task.
- ▶ With all of this: we aim to **drastically reduce the time devoted to text editing**.

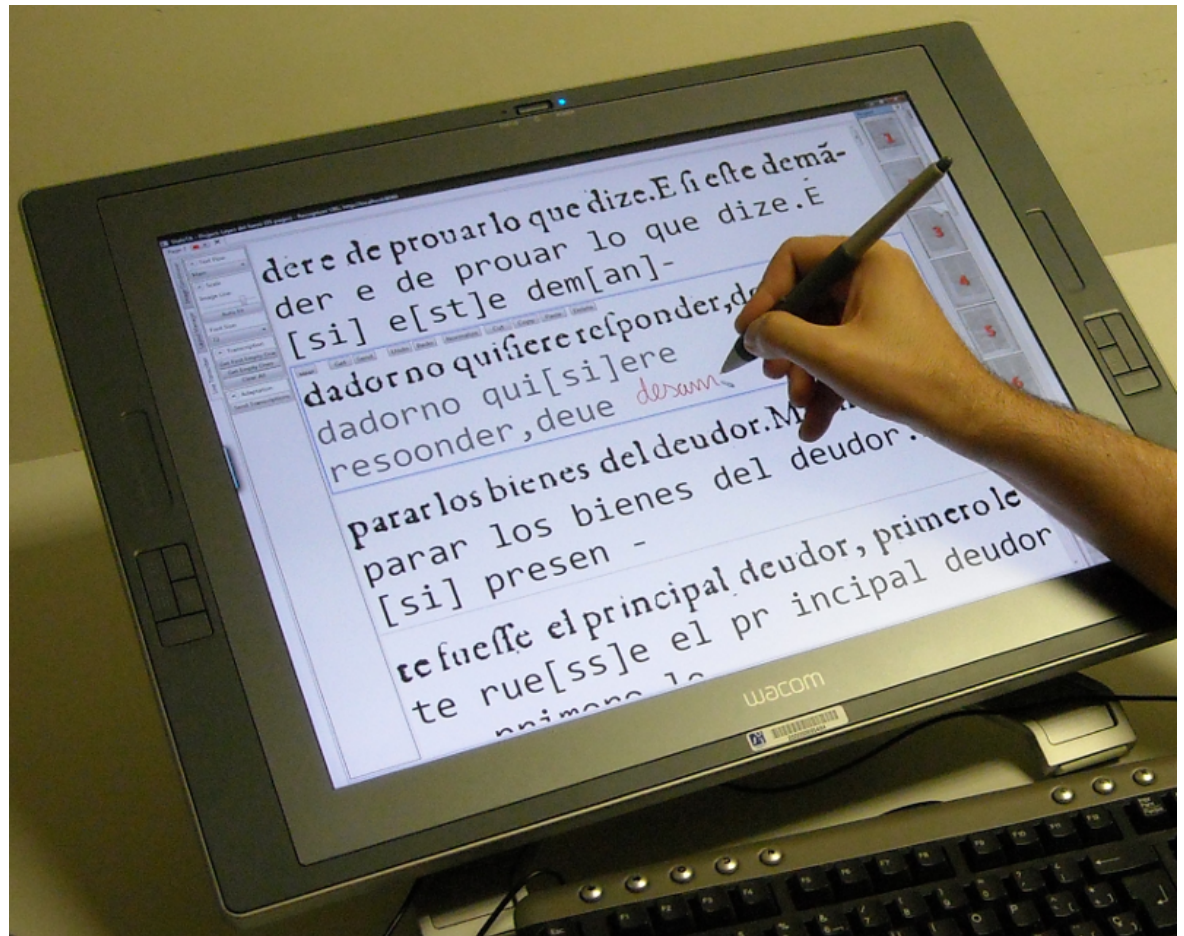
Multimodal interaction

- ▶ **Mouse, keyboard, and stylus.**
- ▶ **Voice**

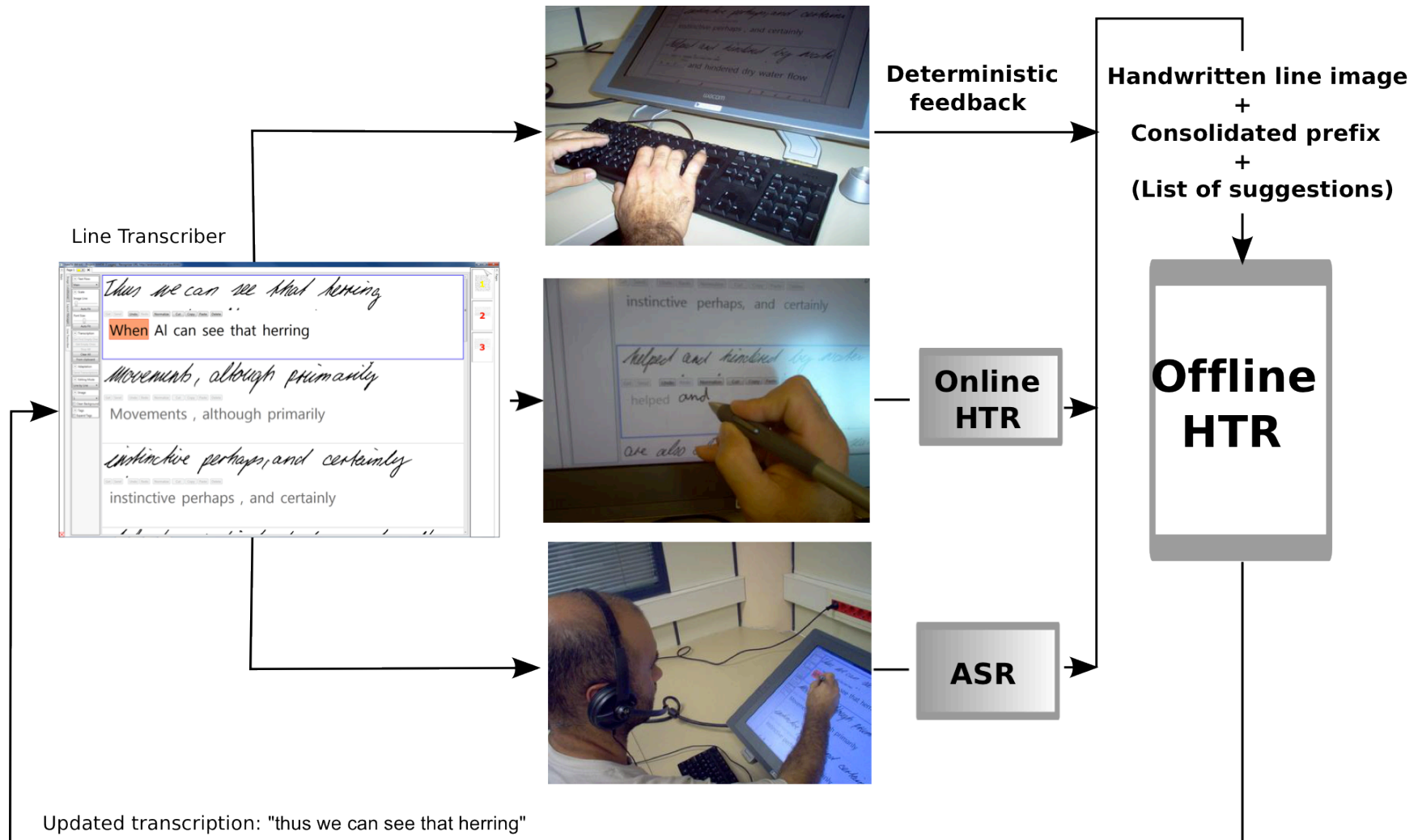


STATE: An Assisted Text Transcription System

- ✓ It allows **to edit text** line by line
- ✓ Comfortable **interaction**
- ✓ **Keyboard, stylus** or **speech**



User-interaction cycle with STATE via keyboard, stylus, or voice



Questions?

